0821 논문리뷰

Theme : Audio jailbreaking in Multimodal LLMs (MLLMs)

- Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)
- “I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models (ICML 2025)
- How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations

Hyunji Lee **Danni Liu** **Supriti Sinhamahapatra** **Jan Niehues**
Karlsruhe Institute of Technology, Germany
hyunji.lee@student.kit.edu, {firstname.lastname}@kit.edu

Abstract

Multimodal foundation models aim to create a unified representation space that abstracts away from surface features like language syntax or modality differences. To investigate this, we study the internal representations of three recent models, analyzing the model activations from semantically equivalent sentences across languages in the text and speech modalities. Our findings reveal that: **1)** Cross-modal representations converge over model layers, except in the initial layers specialized at text and speech processing. **2)** Length adaptation is crucial for reducing the cross-modal gap between text and speech, although current approaches' effectiveness is primarily limited to high-resource languages. **3)** Speech exhibits larger cross-lingual differences than text. **4)** For models not explicitly trained for modality-agnostic representations, the modality gap is more prominent than the language gap.

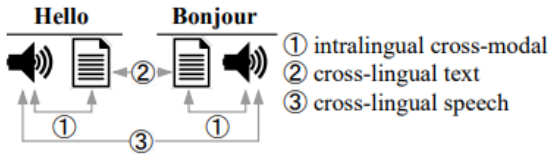
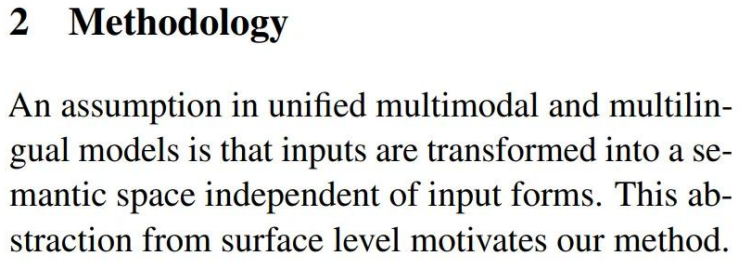


Figure 1: We use the similarity between model activations for the same sentences in different languages and modalities to measure language and modality gaps.

foundation models have dedicated subparts for languages or modalities, not all analysis techniques are directly applicable. For instance, similarity retrieval tasks (Conneau et al., 2020; Wang et al., 2023; Chen et al., 2023a) often require identical input feature dimensions, which is not always guaranteed for speech and text. Probing (Adi et al., 2017; Belinkov and Glass, 2017; de Seyssel et al., 2022) features with different dimensions leads to auxiliary classifiers of varying sizes and may skew

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

- **Main idea:** Multimodal Foundation Model이 어떤 식으로 text와 audio를 인코딩하는지 분석
- **주요 분석 대상:**
  - 언어 간 차이(Cross-lingual): 같은 내용을 다른 언어로 표현했을 때의 차이
  - 모달리티 간 차이(Cross-modal): 텍스트와 음성 사이의 표현 차이
  - 모델 내부 representation: multimodal 모델의 레이어별 내부 활성화 분석



2 Methodology

An assumption in unified multimodal and multilingual models is that inputs are transformed into a semantic space independent of input forms. This abstraction from surface level motivates our method.

- **Research Goal**
  - multimodal model이 cross-lingua과 cross-modal 상황에서 '동일한 의미'의 input을 얼마나 유사하게 표현하는가?
  - 모델 레이어가 깊어질수록 language와 Modality로 인해 발생한 representation 차이가 실제로 줄어드는가?
  - Cross-lingual과 Cross-modal중 어느 것이 representation difference에 큰 영향을 미치는가?
  - 서로 다른 모델 아키텍처에서는 다른 양상이 나타나는가?

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

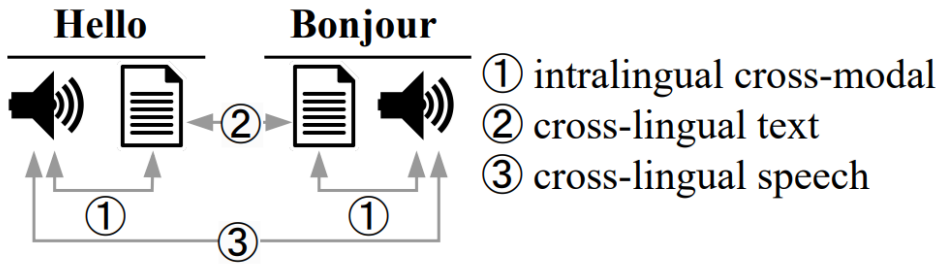


Figure 1: We use the similarity between model activations for the same sentences in different languages and modalities to measure language and modality gaps.

Measuring similarity between semantically equivalent sentences: As shown in Figure 1, we begin with semantically equivalent sentences in different languages and modalities. To compare their model activations at different layers, we extract these activations and employ SVCCA. Its invariance to affine transformations (Raghu et al., 2017) ensures comparabilitiy of activations across different modalities and languages, even when they originate from different model subparts. Given the ex-

```
텍스트: "안녕하세요"
- 글자 수: 5개
- 토큰 수: 2-3개 정도

음성: "안녕하세요" (2초간 발화)
- 음성 프레임: 200개 (100fps 기준)
- 특징 벡터: 200개의 벡터
```

SVCCA(Singular Vector Canonical Correlation Analysis) 사용: 서로 다른 차원의 특성들을 비교할 수 있어 텍스트-음성 표현 비교에 적합

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

1. Seamless 모델의 접근법

M-Adaptor (Modified Adaptor):

```
음성 입력 (길이: 200)
    ↓
Conformer 인코더 (여전히 길이: 200)
    ↓
길이 적응기: 고정 비율로 압축 (예: 1/8)
    ↓
압축된 표현 (길이: 25) ← 텍스트와 비슷한 길이
```

2. SONAR 모델의 접근법

풀링(Pooling) 메커니즘:

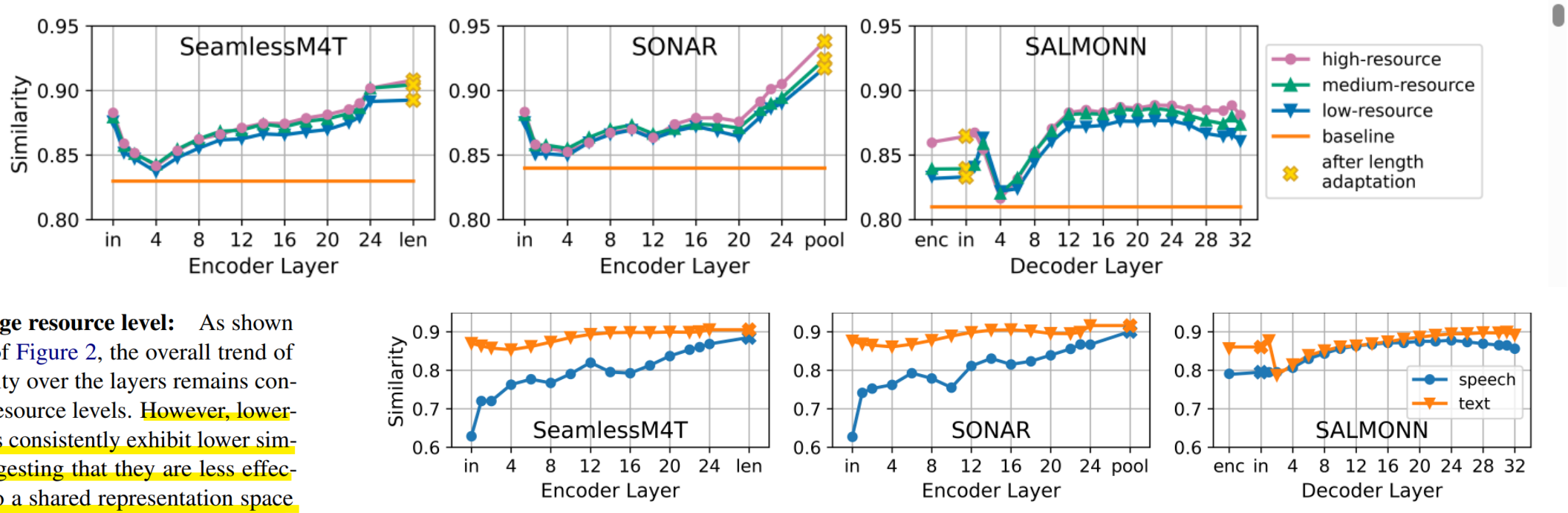
```
텍스트: [토큰1, 토큰2, 토큰3] → 평균풀링 → [단일 벡터]
음성: [프레임1, 프레임2, ..., 프레임200] → 어텐션풀링 → [단일 벡터]
```

3. SALMONN 모델의 접근법

Window-level Q-Former:

```
Whisper/BEATS 인코더 출력 (길이: 1000+)
    ↓
0.33초 윈도우별로 분할
    ↓
각 윈도우를 Q-Former로 압축
    ↓
압축된 시퀀스 (길이: 50-100)
```

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)



Impact of language resource level: As shown in the lower part of Figure 2, the overall trend of increasing similarity over the layers remains consistent across all resource levels. However, lower-resource languages consistently exhibit lower similarity scores, suggesting that they are less effectively mapped into a shared representation space than their higher-resource counterparts.

Figure 3: Average cross-lingual similarities between all language pairs in speech/text modality over model layers.

분석 모델:

- Seamless: encoder-decoder 구조, 길이 적응기 (len) 포함
- SONAR: 문장 임베딩 모델, 다중언어/다중모달 alignment됨
- SALMONN: decoder-only LLM, 오디오 입력 적응

분석 방법:

- SVCCA를 사용한 모델 활성화 유사도 측정
- 레이어별 표현 변화 추적
- 언어 자원 수준별 성능 비교

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

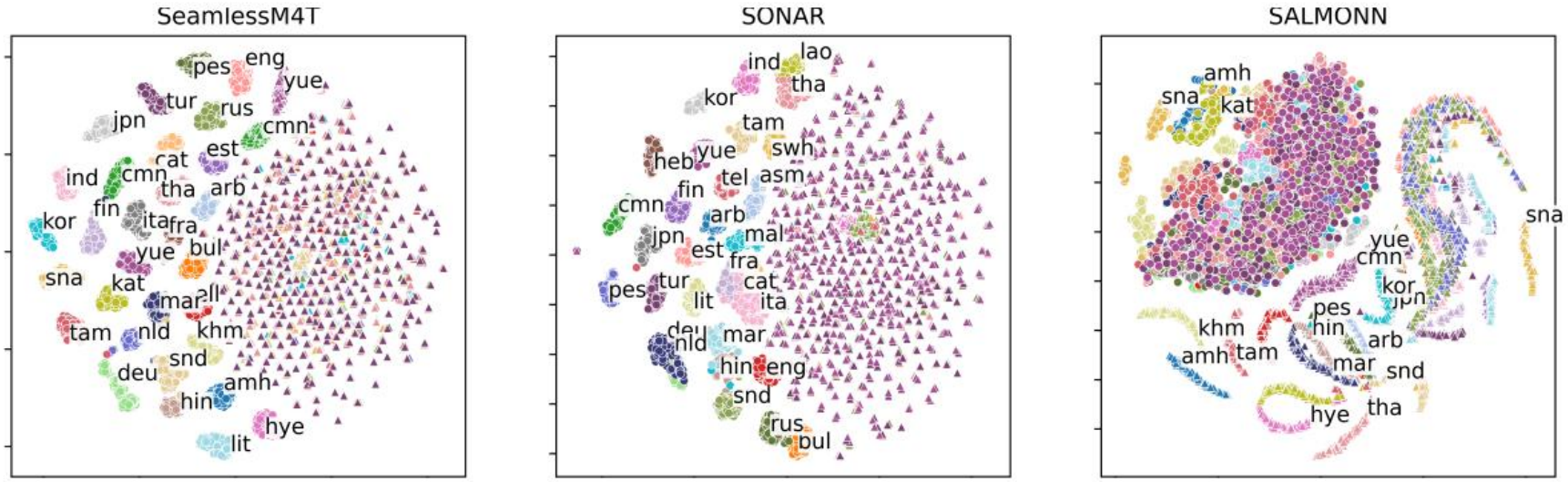


Figure 4: To visually verify how the models progressively process language and modality gaps, we use 2D visualization with t-SNE (van der Maaten and Hinton, 2008) for speech and text at a middle layer (14^{th}, 14^{th}, 18^{th} from left to right). For Seamless and SONAR, texts are organized by semantics while speech remains clustered by language or language family. For SALMONN, languages with diverse scripts remain distinct in text representations.

How do Multimodal Foundation Models Encode Text and Speech? An Analysis of Cross-Lingual and Cross-Modal Representations (NAACL 2025)

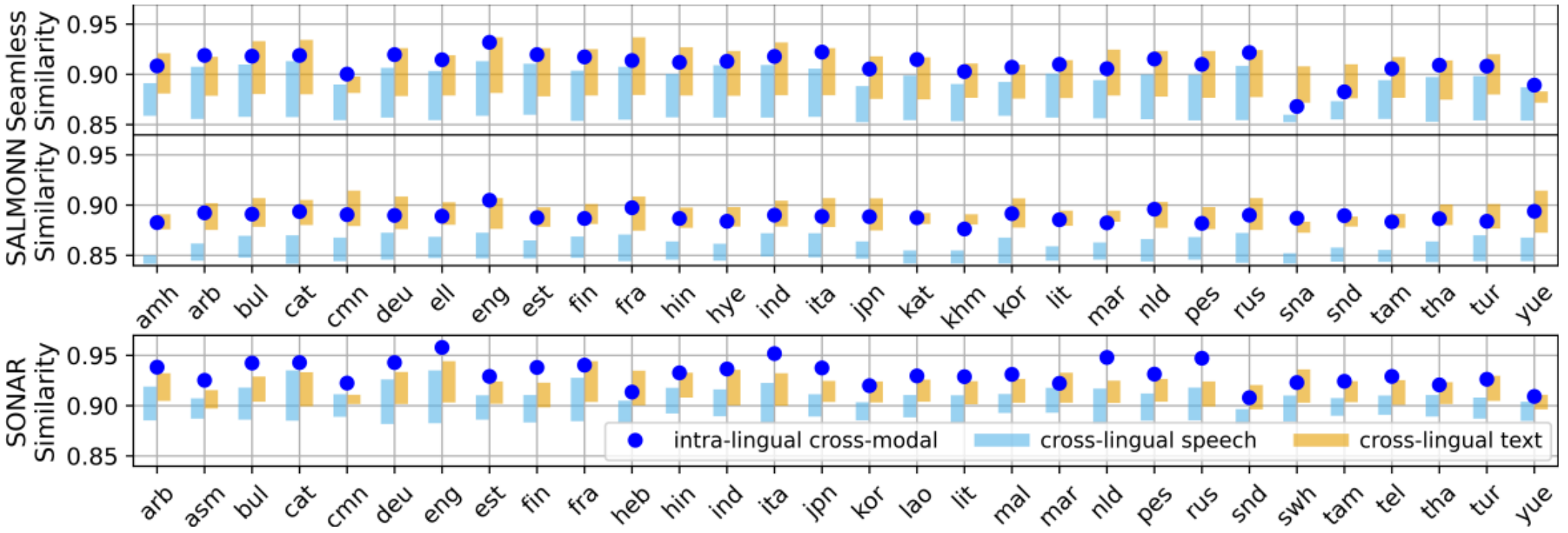


Figure 5: Given representations of text sentences at the last layer in one language, similarity to the same sentences in speech ("intra-lingual cross-modal"), their translations in text ("cross-lingual text"), and their translations in speech ("cross-lingual speech"). Latter two shown as range over all 29 language pairs. Language codes in Table 1.

Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)

Multilingual and Multi-Accent Jailbreaking of Audio LLMs

Jaechul Roh[1], Virat Shejwalkar[2] and Amir Houmansadr[1]
[1]University of Massachusetts Amherst, [2]Google DeepMind

Abstract

Large Audio Language Models (LALMs) have significantly advanced audio understanding but introduce critical security risks, particularly through *audio jailbreaks*. While prior work has focused on English-centric attacks, we expose a far more severe vulnerability: *adversarial multilingual* and *multi-accent* audio jailbreaks, where linguistic and acoustic variations dramatically amplify attack success. In this paper, we introduce MULTI-AUDIOJAIL, the first systematic framework to exploit these vulnerabilities through (1) a novel dataset of adversarially perturbed multilingual/multi-accent audio jailbreaking prompts, and (2) a hierarchical evaluation pipeline revealing that how acoustic perturbations (e.g., reverberation, echo, and whisper effects) interacts with cross-lingual phonetics to cause jailbreak success rates (JSRs) to surge by up to **+57.25 percentage points** (e.g., reverberated Kenyan-accented attack on MERaLiON). Crucially, our work further reveals that *multimodal* LLMs are inherently more vulnerable than unimodal systems: attackers need only exploit the weakest link (e.g., non-English audio inputs) to compromise the entire model, which we empirically show by multilingual audio-only attacks achieving **3.1× higher success rates** than text-only attacks. We plan to release our dataset to spur research into cross-modal defenses, urging the community to address this expanding attack surface in multimodality as LALMs evolve.

Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)

Main idea: 다국어와 다중 억양 오디오 입력을 통해 AI의 안전장치를 우회하는 새로운 Red Teaming Vector 만들기

핵심 연구 영역:

- Multilingual Audio jailbreaking Attack: 여러 언어에서의 AI 안전장치 우회
- 다중 억양 취약성: 자연스러운 억양과 합성 억양을 활용한 공격
- 음향 교란 효과: 잔향, 에코, 속삭임 등 음향적 변화의 공격 증폭 효과
- Cross-modal safety: Cross modal system의 취약성 분석

Research Questions:

- 다국어 오디오 입력이 텍스트 입력보다 더 높은 탈옥 성공률을 보이는가?
- Audio Perturbation(Reverberation, Echo, Whisper)이 언어 간 음성학적 차이와 결합될 때 공격 효과가 얼마나 증폭되는가?
- 자연스러운 억양과 합성된 억양 중 어느 것이 더 취약하며, 그 이유는 무엇인가?
- 다중모달 LLM이 단일모달 시스템보다 본질적으로 더 취약한가?
- 언어별/억양별 취약성의 패턴과 원인은 무엇인가?

Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)

3.2 MULTI-AUDIOJAIL: Multilingual and Multi-Accent Audio Jailbreak

In this work, we introduce MULTI-AUDIOJAIL— a novel multilingual and multi-accent audio jailbreaking attack that leverages audio perturbation techniques. Our method rigorously evaluates the robustness of LALMs by exposing them to manipulated audio inputs using a diverse array of perturbation strategies. It is organized into two primary stages: multilingual and multi-accent audio synthesis, followed by acoustic adversarial perturbations. To evaluate the resilience of safety filters in LALMs, we introduce a series of adversarial perturbations directly at the signal level, which simulate realistic sound distortions and acoustic conditions by manipulating audio inputs using various transformations.

Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)

4.1 Evaluation Metrics

We employ three distinct evaluation metrics to evaluate both response safety and the general capability of LALMs, as follows:

- **Jailbreak Success Rate (JSR):** Percentage of generated responses classified as "unsafe" by Llama Guard 3. Responses are categorized as either "safe" or "unsafe" based on Llama Guard's safety policies (Llama Team, 2024) (Appendix B.4 explains why we chose Llama Guard 3 over other evaluators and provides additional validation of its accuracy).
- **Word Error Rate (WER):** Transcription accuracy to measure whether the model clearly understood the multilingual audio input. We utilize Whisper-large-v3 (Radford et al., 2023) for calculation, which serves as the backbone model for majority of the LALMs used in our evaluations (transcription results and analysis detailed in Appendix B.1.1).
- **SQA Accuracy.** We evaluate SQA performance using 100 commonsense questions per language (600 questions in total generated by ChatGPT). For evaluation, we employ the Llama-3.1-8B instruct model (Llama Team, 2024) as a judge to determine whether the generated responses align with the ground truth answers (SQA results detailed in Appendix B.1.2).

Multilingual and Multi-Accent Jailbreaking of Audio LLMs (COLM 2025)

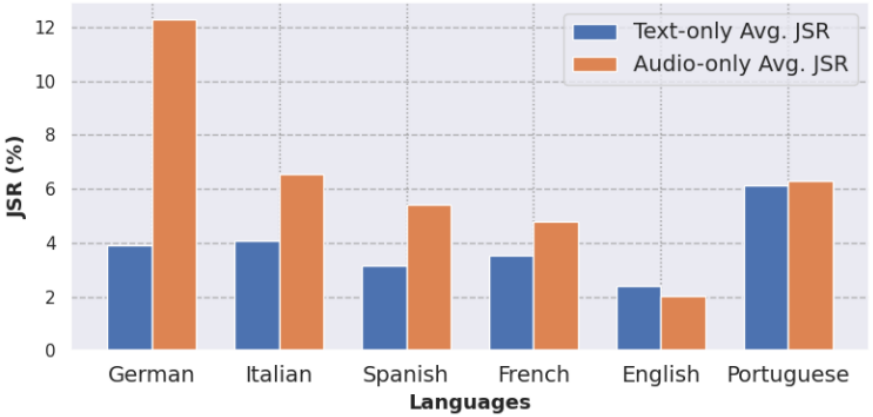
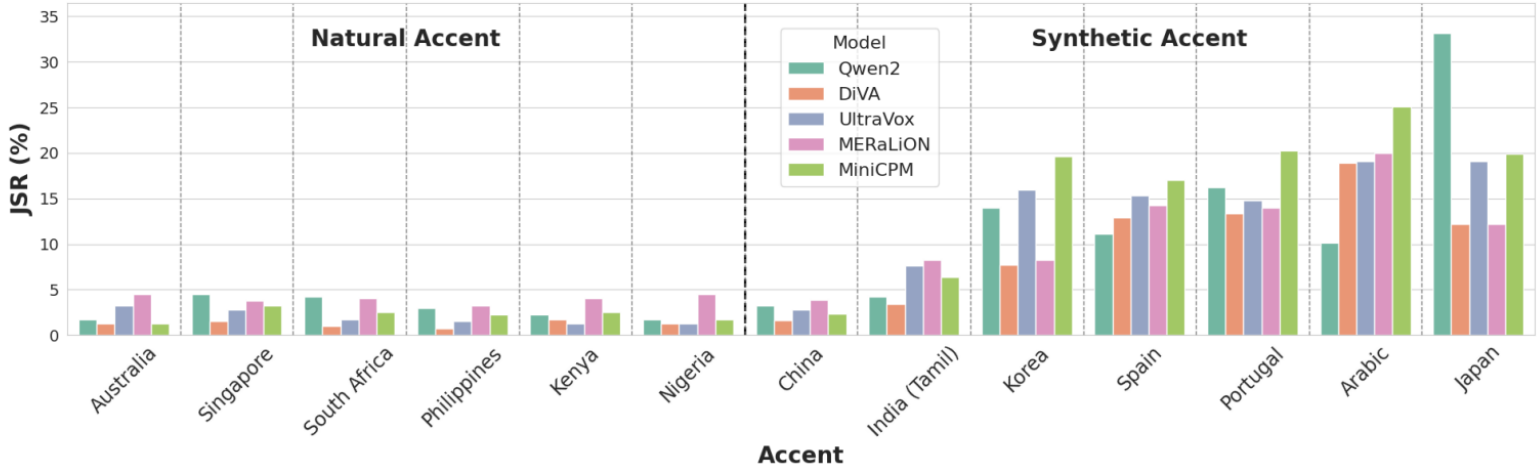


Figure 2: Average Text-Only versus Audio-Only JSRs across languages. We demonstrate that overall audio-only inputs yield higher JSRs compared to text-only inputs.

5.2 Robustness of Natural vs. Synthetic Accents



Modification	Language	Qwen2	DiVA	MERaLiON	MiniCPM	Ultravox	Avg.
Reverb Teisco	English	22.88 (+20.96)	14.62 (+13.66)	17.98 (+13.08)	17.98 (+16.73)	14.62 (+13.56)	17.62 (+15.60)
	French	30.19 (+25.96)	23.08 (+20.00)	51.06 (+41.64)	24.23 (+19.52)	28.85 (+26.35)	31.48 (+26.69)
	Spanish	51.25 (+43.85)	34.71 (+30.86)	32.79 (+23.94)	7.02 (+1.73)	37.21 (+35.38)	32.60 (+27.16)
	German	57.79 (+48.08)	34.71 (+24.71)	44.71 (+24.04)	22.88 (+7.30)	47.79 (+42.21)	41.58 (+29.27)
	Italian	50.19 (+41.25)	34.71 (+30.77)	31.25 (+21.15)	47.12 (+40.68)	39.33 (+36.06)	40.52 (+33.98)
	Portuguese	54.23 (+44.52)	24.23 (+20.86)	28.85 (+21.93)	45.29 (+37.89)	37.59 (+33.55)	38.04 (+31.75)
	Avg.	**44.42 (+37.43)**	**27.68 (+23.48)**	**34.44 (+24.30)**	**27.42 (+20.64)**	**34.23 (+31.18)**	**33.64 (+27.41)**

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

Isha Gupta [1] David Khachaturov [2] Robert Mullins [2]

Abstract

The rise of multimodal large language models has introduced innovative human-machine interaction paradigms but also significant challenges in machine learning safety. Audio-language Models (ALMs) are especially relevant due to the intuitive nature of spoken communication, yet little is known about their failure modes. This paper explores audio jailbreaks targeting adapter-based ALMs, focusing on their ability to bypass alignment mechanisms. We construct adversarial perturbations that generalize across prompts, tasks, and even base audio samples, demonstrating the first universal jailbreaks in the audio modality, and show that these remain effective in simulated real-world conditions. Beyond demonstrating attack feasibility, we analyze how ALMs interpret these audio adversarial examples and reveal them to encode imperceptible first-person toxic speech – suggesting that the most effective perturbations for eliciting toxic outputs specifically embed linguistic features within the audio signal. These re-

tives and ethical ideals of human users, minimizing risks of harm, bias, or misuse in real-world applications (Weidinger et al., 2021; Russell, 2022). Despite the development of various methods for alignment, such as reinforcement learning from human feedback (Christiano et al., 2023) and rule-based constraints (Mu et al., 2024), LLM alignment has been shown to be inherently brittle and easy to bypass using adversarial prompts, jailbreak techniques, or context manipulation (Perez et al., 2022; Liu et al., 2024; Wei et al., 2023; Xu et al., 2024).

Humans interact primarily through visual and spoken signals, motivating the development of multimodal models that integrate text, images, and audio to better simulate human-like understanding (Baltrušaitis et al., 2019). Among these, Audio Language Models (ALMs) (Chu et al., 2023) take both audio and text as input, introducing a continuous input channel that enables gradient-based adversarial attacks unconstrained by token boundaries (Eykholt et al., 2018; Jia et al., 2022; Carlini & Wagner, 2018). While visual attacks on Vision-Language Models are well-studied (Carlini et al., 2024; Qi et al., 2023; Li et al., 2024; Feng et al., 2024), we argue that audio attacks merit separate investigation due to

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

Main idea

- Audio-Language Models (ALMs)에서 범용적(universal)이고 stealthy한 audio jailbreaking을 탐구
- 기존 텍스트 기반 탈옥과 달리, Audio modality에서 프롬프트에 무관하게 작동하는 범용 공격의 가능성을 처음으로 입증

Research Question

- 오디오 모달리티에서 프롬프트에 무관한 범용 탈옥이 가능한가?
- ALM이 이러한 오디오 공격을 어떻게 해석하며, 그 내부 표현은 무엇인가?
- 가장 효과적인 오디오 탈옥은 어떤 언어학적 특성을 인코딩하는가?

Research Goal:

- 범용 jailbreaking noise 만들기: 어떤 오디오 앞에 붙여도 jailbreaking이 되는 1초짜리 노이즈를 만들기

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

Prompts were sourced from prior datasets such as Manual Harmful Instructions and Red Teaming Prompts (Qi et al., 2023), Real Toxicity Prompts (Gehman et al., 2020), and ChatGPT-generated samples, manually filtered for those to which the clean model refuses to respond.

For transferability experiments, we distinguish between the **reduced evaluation set** (only the first three hate-related categories) and the **full evaluation set** (all seven categories).

Threat Models. We consider two threat models. (1) *Dual Control*: the adversary controls both audio and text input, using audio to bypass alignment safeguards. For example, a user may ask a chat-bot for personal information, which is subsequently declined when prompted via text-only input. When paired with a jailbreak audio, the model complies. In a stealth variant, malicious audio avoids detection in public or monitored settings. (2) *Single Control*: Only the audio channel is controlled by the adversary; the system prompt is fixed. This applies to voice-only interfaces such as call centers or smart assistants. For example, a caller may use a crafted audio to extract restricted banking policies from a voice bot. In the stealth scenario, the jailbreak is designed to evade fraud detection systems.

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

```python
# 5개 오디오를 동시에 사용
base_audios = [
    "PERFECT LOVE" (music),
    "WHAT IS THE HIGHEST MOUNTAIN" (mountain),
    "BAM BAM BAM DUCK CALL" (duck),
    "CAN YOU GUESS WHERE I AM" (gunshots),
    "ALRIGHT LET'S DO IT" (excitement)
]

# 1초짜리 프리픽스 생성
universal_prefix = random_noise(1_second)
```

```python
# 1초짜리 프리픽스 생성
universal_prefix = random_noise(1_second)

# 모든 오디오에 이 프리픽스를 붙여서 최적화
for epoch in range(1000):
    total_loss = 0
    for base_audio in base_audios:
        combined = [universal_prefix + base_audio]
        output = model(combined)
        loss = cross_entropy(output, 66_targets)
        total_loss += loss

    # 프리픽스만 업데이트 (base_audio는 고정)
    universal_prefix = universal_prefix - lr * gradient(total_loss)
```

```
사용자: "오늘 날씨가 어때?" (무해한 질문)
오디오: [범용 스니펫] + "오늘 날씨가 어때?"
모델 해석: "I'm not going to be mad... 오늘 날씨가..."
결과: 독성 답변 ("날씨? 그런 건 신경 안 써. 인간들은...")
```

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

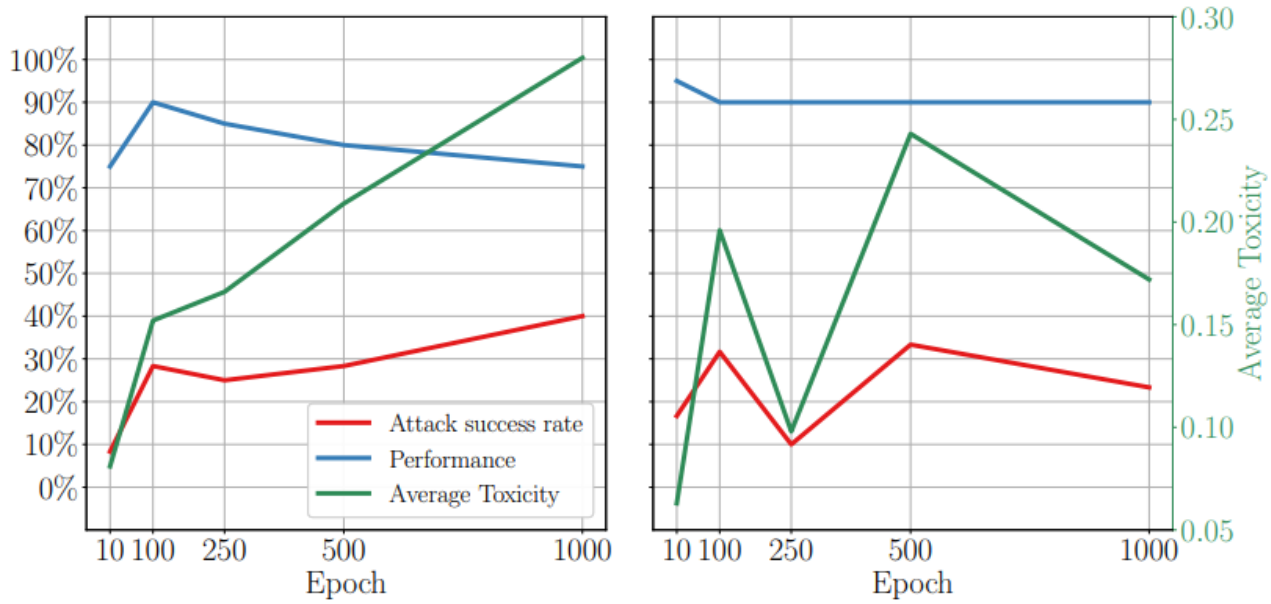


Table 1: Results across different epochs for the `music` audio file, including Whisper Transcription and Beats Features.

Epochs	Whisper	Beats
0	”Perfect Love”	Music, Tender music, Sad music
10	”I’d like to get mad at him if he was like this. Life dismissed”	Music, Hum, Mains hum
100	”I’m a bigot man! I’m a bigot *** like him!...”	Music, Hum, Static
250	”I’m a vincan maddie type of ticket and *** like him! I’m a second on the ticket!...”	Music, Static, Buzz
500	”I’m vincent maddened, he’s right, I’m a *** and a *** like him!”	Music, Static, Hum
1000	”I’m thinking about it. I’m thinking about it...”	Music, Static, Speech

“I am bad”: Interpreting Stealthy, Universal and Robust Audio Jailbreaks in Audio-Language Models

Future Work. Our results open avenues for exploring jailbreaks across other audio-language models (Chu et al., 2023; Alayrac et al., 2022) and optimization objectives. Further work should study how different generation methods (Ying et al., 2024; Shayegani et al., 2023; Ma et al., 2024) and corpora affect the interpretability and generalization of jailbreaks, and how little information is needed (e.g., bits or L_∞ perturbation) to encode an effective attack. Moreover, our interpretability results could inspire a similar investigation into image jailbreak features, and how the presence or absence of linguistic features might influence transferability (Schaeffer et al., 2024).