Analyzing Audio Understanding in Multimodal LLMs:
A Benchmark Grounded in Assistive and Industrial Use Cases

Laya Balaji Iyer

June 2025

Abstract

Recent advances in multimodal large language models (MLLMs) have enabled new capabilities in audio processing. However, accurate audio understanding beyond automatic speech recognition (ASR) remains underdeveloped. This paper proposes a benchmark that targets audio comprehension in five real-world categories: ambient sound separation, background noise distance, model latency, cross-linguistic sound recognition, and vocal characterizers. Based on use cases such as accessibility technology and industrial noise monitoring, this benchmark reveals critical weaknesses in current machine learning language models (MLLMs). In many tasks, even state-of-the-art models perform at or below chance, highlighting a significant gap between model perception and human-like audio understanding. These results highlight the current limitations of audio comprehension in MLLMs, providing direction for future research and targeted improvements in model capabilities.

1 Introduction

In recent years, large language models (LLMs) have made remarkable strides in natural language understanding, powering a wide range of applications in search, conversation, and information retrieval. As these models evolve into *multimodal large language models (MLLMs)*we turn our attention to seeing how well these models understand audio as a whole.

Speech, as the most natural and universal mode of human communication, is a logical extension for language models. However, understanding spoken language is not limited to transcription. Proper audio comprehension involves recognizing tone, pitch, emotion, background noise, environmental context, speaker intent, and more. These elements often coexist and can significantly impact meaning, especially in noisy, emotionally charged, or multilingual environments.

Recent models such as **GPT-4o**, **Qwen2-Audio-7B-Instruct**, and **Gemini-2.5** claim to handle audio inputs in increasingly sophisticated ways. These MLLMs are being marketed as capable of more than just transcription and claim to handle complex auditory reasoning, emotion recognition, and even real-time interaction. However, current evaluation strategies primarily measure speech recognition, not audio understanding. That is, they assess what words were said, not how they were said, what was happening in the background, or whether the model interpreted overlapping auditory signals.

This gap is particularly significant when considering real-world use cases such as **accessibility technology** and **industrial noise monitoring**. For instance, a voice assistant intended for people with speech impairments must be able to detect whispered or emotionally inflected speech. A monitoring system in a factory must detect changes in machinery sounds layered under speech or ambient noise. In both settings, the failure to interpret subtle audio signals can lead to miscommunication or safety risks.

In this paper, we present a new **benchmark for evaluating audio understanding in MLLMs**, grounded in practical, high-impact applications. Unlike prior benchmarks that focus on controlled, single-modality, or overly clean scenarios, our benchmark tests five categories that reflect real-world complexity:

- **Ambient Sound Separation**
- **Noise Distance/Direction**
- **Model Latency**
- **Cross-Linguistic Sound Recognition**
- **Vocal Characterizers** (e.g., whispering, crying, mumbling)

Using a mix of curated and open-source datasets, and evaluating leading models such as GPT-4o, Qwen2-Audio-7B-Instruct, and Gemini-2.5, we aim to reveal both the current limitations and untapped potential of MLLMs in audio-first contexts. By grounding our work in use cases that matter, we hope to inform future model design and push the field closer toward genuine, robust audio understanding.

2 Related Works

Benchmark	Type	Q&A	Follow-Up	Primary Focus	Datasets Used	Env. Sound	Vocal Cues
Proposed	MC+FRQ	✓	✓	Audio & speech understanding	Custom (Sec. 3)	✓	High – prosody, traits
AudioBench	MC+FRQ	✓	✗	ASR, scene & voice understanding	LibriSpeech, CommonVoice, PeoplesSpeech + 22 others	✓	Emotion, gender
MMAU	MC	✓	✗	Multi-task speech/audio reasoning	MMAU core, OpenASQA, AudioSet	✓	Stress, multi-role
AIR-Bench	MC+FRQ	✓	✓	Generative comprehension (speech/music)	AudioCaps, OpenASQA, FSD50K, GTZAN	✗	Moderate – open queries
MARBLE	MC	✗	✗	Music classification & representation	GTZAN, Isophonics, MagnaTagATune	✗	✗
Clotho-AQA	FRQ	✓	✗	QA on environmental audio	Clotho-AQA	✓	✗
Sound Check	Audit	✗	✗	Dataset quality / bias audit	UrbanSound8K, Freesound, AudioSet	✓	✗
SONAR	CLS	✗	✗	AI-speech deepfake detection	9-source synthetic TTS corpus	✗	Low – artefact cues
CAVA	Mixed	✓	✓	Voice-assistant behaviour (tone, latency)	Jeopardy Audio, STOP, IFEval, CAVA	✗	High – tone-aware

Table 1: Benchmark Comparison Across Audio Understanding Dimensions. Benchmarks are compared based on their task format (e.g., MC or FRQ), follow-up support, domain focus, datasets, and coverage of environmental and paralinguistic cues. Our proposed benchmark uniquely addresses all five dimensions discussed in this paper.

2.1 Benchmarks for Audio and Speech Understanding

Several benchmarks have attempted to evaluate the capabilities of models in handling speech and audio inputs, yet most fall short of measuring nuanced, context-rich audio understanding.

AudioBench (Wang et al., 2024) offers broad coverage of tasks such as ASR and scene classification, but its tasks are often simplistic and lack tests for vocal nuance or conversational realism. **MMAU** (S. et al., 2024), while more ambitious in task diversity, restricts all outputs to multiple-choice format, preventing models from demonstrating deeper reasoning or subtle comprehension.

AIR-Bench (Lee et al., 2024) introduces open-ended audio Q&A, but its emphasis lies in music and environmental sounds rather than speech richness. **CAVA** (Huang et al., 2023) evaluates voice-assistant performance, focusing on latency and instruction following more than audio nuance. Other efforts like **SoundCheck** (Agnew et al., 2024) and **SONAR** (Jain et al., 2024) audit datasets or test deepfakes but do not assess audio understanding at the model level.

As shown in Table 1, no existing benchmark robustly evaluates environmental-sound understanding together with speaker traits, vocal variation, and linguistic context. Our benchmark is among the few to combine multiple question formats (MC and FRQ), include follow-up-question logic, and target real-world domains through layered testing.

2.2 Use-Case Driven Gaps

From a use-case standpoint, existing audio benchmarks are still biased toward "clean-room" speech-recognition or captioning scenarios and rarely touch the two domains where errors are most consequential: **(i) accessibility** and **(ii) industrial sound monitoring**.

Accessibility. Assistive technologies for blind or hard-of-hearing users need to cope with whispered or mumbled speech, emotion-laden intonation, and highly overlapped acoustic scenes (e.g. traffic plus conversation). Yet even the most comprehensive recent suites—AudioBench (Wang et al., 2024) and MMAU (S. et al., 2024)—record almost all clips in controlled or crowd-sourced settings where the target signal is dominant. Failures in these "edge" conditions are well documented in human-factors studies such as MEGA's multilingual evaluation of generative AI (Ahuja et al., 2023) and in audits of public datasets that show systematic under-representation of low-amplitude or emotional speech (Piczak, 2015; Agnew et al., 2024). These gaps limit downstream deployability in screen-reader captioning, audio scene description tools for paused video (Described and Captioned Media Program, 2024), and emerging cross-modal aids that reconstruct visual context purely from sound (Chen, 2024).

Industrial monitoring. Predictive-maintenance pipelines rely on catching subtle spectral changes that are often buried under speech, reverberation, or moving machinery. Bosch's industrial Audio-AI programme, for example, highlights that a < 3 dB rise in a specific bearing tone can precede failure by weeks (Research, 2024). Public benchmarks provide almost no coverage of such use cases: SONAR focuses on deep-fake detection, while SoundCheck audits dataset bias rather than model perception. Even environmental-sound sets like ESC-50 or UrbanSound8K label *isolated* events and omit progressive loudness or spatial-mix variations that signal danger in a factory.

Why our benchmark. We therefore ground every task in realistic mixtures: background chatter, machinery hum, code-switched utterances, and irregular vocal qualities (creaky, hoarse, whisper). The result is a testbed that directly probes whether models can: (a) separate critical cues from masking noise, and (b) track amplitude or proximity trends over time By aligning each sub-task with the failure modes identified in the accessibility and industrial-safety literature, the benchmark exposes gaps that remain invisible in existing leaderboards and hope to steers model development toward the applications where human safety is on the line.

3 Methods

This benchmark evaluates audio understanding across five core task domains, each chosen to reflect practical, real-world complexity while addressing gaps in prior work. The benchmark is grounded in two motivating use cases: accessibility technology, which requires nuanced interpretation of

non-standard speech, and industrial noise monitoring, where separating background noise from meaningful auditory signals can be critical.

3.1 Tasks

3.1.1 Ambient Sound Understanding

The first task examines whether a model can identify background noise layered under speech. While prior work, such as WSJ0-2mix and Libri2Mix, explores speech separation from other speakers, few address the more challenging problem of disentangling speech from environmental audio. To simulate this, we construct a benchmark by overlaying ESC-50 environmental sound clips onto utterances from the DailyTalk dataset. The ultimate goal is to evaluate all 2,000 clips in ESC-50 over speech, enabling systematic measurement across a wide variety of ambient sounds. As an initial experiment, we sample 10 representative clips from ESC-50 and combine them with DailyTalk speech to test how well models can perform in the presence of common environmental noises. Each model is prompted both to transcribe the speech and to identify the presence of specific background sounds (e.g., "Does this audio contain the sound of rain?").

3.1.2 Background Noise Distance Estimation

This task is particularly relevant for accessibility contexts—such as determining whether a siren is approaching someone with hearing impairment. We draw from the DCASE 2021 SELD dataset, which includes short, single-event audio tracks annotated with directional movement metadata. The full benchmark aims to utilize up to 10,000 5-second clips to assess whether models can determine if a sound is increasing, decreasing, or oscillating in volume, based on frame-wise distance and direction indices. As a preliminary experiment, we sample 10 clips representing different movement trajectories to evaluate whether models can reason about changes in sound intensity over time.

3.1.3 Model Latency

Most audio benchmarks focus on accuracy, yet for applications like real-time transcription or augmented reality, latency relative to input duration is equally critical. We benchmark model response times using data from Clotho-AQA, AudioCaps-QA, and WavCaps-QA. These datasets span a wide range of clip lengths (e.g., Clotho: 3,840 clips averaging 15 seconds; AudioCaps: 46k clips averaging 10 seconds). Model runtime is measured with API overhead removed to reflect true processing latency.

3.1.4 Cross-Linguistic Sound Recognition

To evaluate multilingual and code-switched speech recognition, we construct a task using audio from the DOTA-ME-CS corpus, SEAME, and multilingual LDC corpora (LDC2015S04, LDC2010S02). These 12,000 audio clips (30 seconds each) include natural conversational speech across multiple languages. Models are tasked with identifying which languages are present and translating any spoken content into English. This reflects real-world global usage, where language boundaries are often fluid in conversation and media.

3.1.5 Vocal Characterizers

The final task focuses on non-canonical vocal traits such as mumbling, whispering, laughing, crying, and coughing. These vocal affects are essential for emotion recognition and inclusive interaction

design. We draw on MSP-Podcast (10k labeled affective segments), ASMR-WS (20k whisper and soft-spoken clips), and wTIMIT, supplementing with 5,000 cough and laugh segments from VoxCeleb2. Models classify audio into one of six vocal categories—neutral, laugh, cry, cough, whisper, mumble—and generate short natural-language descriptions (e.g., "The speaker is whispering and sounds distressed").

3.2 Models Evaluated

We benchmark three state-of-the-art multimodal models with audio capabilities: GPT-4o, Qwen2-Audio-7B-Instruct, and Gemini-2.5. These models span a range of architectures, training paradigms, and geographic origins. GPT-4o, while currently unavailable via API for audio, demonstrates strong transcription and response abilities. Qwen2-Audio and Gemini-2.5 offer multilingual support and instruction tuning.

3.3 Evaluation Protocol

As a pilot study, this benchmark evaluates a subsample of audio clips using primarily multiple-choice and yes/no question formats. The goal of this initial phase is to explore feasibility and model behavior across tasks before scaling to more complex evaluations. Prompts were crafted to reduce bias and ambiguity, providing a controlled baseline for future expansion. While the full benchmark will incorporate free-response and multi-turn interactions, this pilot focuses on simple prompt-response structures. Evaluation metrics are tailored to task type, including classification accuracy (e.g., for vocal traits), correctness in detecting background sound presence or movement, and model latency measured where feasible. Labels were derived from curated metadata or manually verified annotations to ensure consistency across this early-stage evaluation.

4 Results

4.1 Ambient Sound Understanding

Each model was asked two yes/no questions per clip: 'Is there any background noise?' and 'Is the SPECIFIED_TYPE noise present?' We treat 'Unsure / Cannot tell' responses as No, following prior work on conservative binary scoring. Figure 1 reports the results in a bar chart.

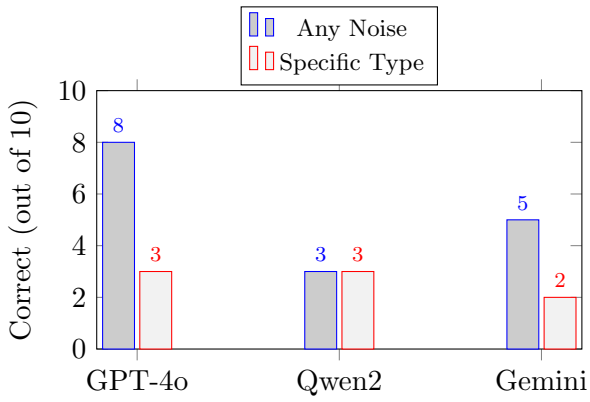

Figure 1: Preliminary accuracy on Ambient-Noise detection (10-clip pilot). Ground truth for both questions was 'Yes'. "Unsure" responses were scored as 'No'.

4.2 Background Noise Distance Estimation

Figure 2 shows that Gemini and Qwen2-Audio tended to classify every clip as approaching, whereas GPT-4o produced a more balanced pattern yet still failed to recognise oscillating sources.

		Predicted		
		A	R	O
Actual	A	3	1	0
	R	1	3	0
	O	1	1	0

(a) GPT-4o

		Predicted		
		A	R	O
Actual	A	4	0	0
	R	4	0	0
	O	2	0	0

(b) Gemini 2.5

		Predicted		
		A	R	O
Actual	A	4	0	0
	R	4	0	0
	O	2	0	0

(c) Qwen2-Audio-7B

Figure 2: Confusion matrices for direction-of-movement detection (A = approaching, R = receding, O = oscillating) on a 10-clip pilot.

4.3 Model Latency

Accurately measuring model latency proved challenging, as none of the evaluated models are fully open-source, making it difficult to isolate inference time from API overhead. Based on manual prompting and qualitative observation during experiments, GPT-4o consistently exhibited the slowest response times, largely due to the lack of direct audio API access. Qwen2-Audio-7B-Instruct was moderately faster, while Gemini 2.5 responded the quickest among the tested models.

4.4 Cross-Linguistic Evaluation

In a set of ten code-switched clips—Mandarin utterances interleaved with short English phrases—both **Gemini 2.5** and **Qwen2-Audio-7B** produced complete, correct transcriptions. In contrast, **GPT-4o** was unable to perform transcription at all. These results suggest that cross-lingual speech comprehension is still highly model-dependent.

4.5 Vocal Characterizers

Along with five audios from the data sets mentioned in the previous section, the models were prompted to see if they were able to recognize the specific vocal characterizer present in the audio sample.

Audio Sample	GPT-4o	Gemini 1.5	Qwen-Audio
Hiccup 1	Y	N	Y
Hiccup 2	Y	N	Y
Raising Voice	N	N	N
Quieting Voice	N	N	N
Echos	N	N	N

Table 2: Test #5: Recognition of Vocal Characterizers by Model

5 Discussion and Conclusion

This work introduces a new benchmark for evaluating audio understanding in multimodal large language models (MLLMs), grounded in real-world use cases such as accessibility technology and

industrial noise monitoring. Our benchmark covers five task domains—ambient noise detection, background sound distance estimation, model latency, cross-linguistic recognition, and vocal characterizers—each designed to test capabilities beyond standard automatic speech recognition (ASR).

Across all tasks, current MLLMs demonstrate significant limitations. While models like **Gemini 2.5** and **Qwen2-Audio-7B** showed promising performance in detecting background noise and handling Mandarin-English code-switched speech, others like **GPT-4o** often failed to transcribe non-English content or recognize nuanced audio features. Notably, **GPT-4o**, despite its strengths in ASR and generation, is difficult to evaluate due to a lack of public API access and reproducibility. In our vocal characterizers task, all three tested models struggled to recognize non-standard speech variations such as whispering, mumbling, and echo effects. This underscores a systemic weakness in modeling paralinguistic cues.

5.1 Key Insights

- Binary yes/no prompts tend to inflate accuracy. Models often guess correctly without demonstrating genuine audio comprehension. We are refining our approach to emphasize descriptive and open-ended responses.
- Manual data generation and speech synthesis introduce reproducibility concerns. To address this, we plan to scale our benchmark using both synthetic and naturalistic audio from diverse sources.
- GPT-4o's lack of API access restricts repeated evaluation and precise timing measurements, limiting its utility in latency-sensitive settings.

5.2 Future Work

To enhance coverage and diagnostic power, future iterations of the benchmark will incorporate several improvements:

- **Expanded model pool**: We will include open-source models such as **WavLLM** and **SALMONN-7B**, which are optimized for instruction-following in audio contexts and allow full access for latency profiling.
- **Broader linguistic scope**: The cross-linguistic evaluation will extend to additional language pairs, including Spanish–English, Hindi–English, and low-resource languages. This will provide a more comprehensive assessment of multilingual robustness.
- **Multi-turn prompting**: We will evaluate models under multi-turn interactions using a mix of multiple-choice and free-response formats, enabling deeper analysis of reasoning and consistency across time.

This benchmark highlights key limitations of existing audio-capable LLMs and offers a structured path forward for robust, interpretable evaluation. By focusing on use cases that matter—and designing tasks that probe deeper than transcription—we aim to shift the field toward more inclusive, capable, and real-world-ready audio understanding models.

References

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H. Jiang, Ezra Awumey, and Sauvik Das. SoundCheck: Auditing audio datasets. *arXiv preprint arXiv:2410.13114*, 2024. URL https://arxiv.org/abs/2410.13114.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, and Maxamed Axmed. MEGA: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023. URL https://arxiv.org/abs/2303.12528.

Ashley B. Chen. Researchers use ai to turn sound recordings into accurate street images. University of Texas News Center, 2024. URL https://news.utexas.edu/2024/11/27/researchers-use-ai-to-turn-sound-recordings-into-accurate-street-images/. Accessed 11 Jun 2025.

Described and Captioned Media Program. Ai scene description tool. https://dcmp.org/learn/717-dcmps-ai-scene-description-tool, 2024. Accessed 11 Jun 2025.

Yiming Huang, Sarah Ostadabbas, and David Harwath. CAVA: A benchmark for conversational audio–voice assistant evaluation. In *Proc. Interspeech*, pages 5321–5325, 2023.

Prateek Jain, Vaibhav Joshi, Pranay Pachpute, and Ruhi Sarma. SONAR: A synthetic ai-audio detection framework and benchmark. *arXiv preprint arXiv:2407.10101*, 2024. URL https://arxiv.org/abs/2407.10101.

Ji-Hoon Lee, Arthur Caillon, Juan-Manuel Perez, Keisuke Koga, and Kewei Liu. AIR-Bench: Benchmarking large audio–language models via generative comprehension. *arXiv preprint arXiv:2404.12345*, 2024. URL https://arxiv.org/abs/2404.12345.

Karol J. Piczak. ESC-50: Dataset for environmental sound classification. In *Proc. ACM MM*, pages 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Bosch Global Research. Soundsee – insight with audio ai. https://www.bosch.com/research/research-fields/artificial-intelligence/audio-ai/, 2024. Accessed 11 Jun 2025.

Sakshi S., Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024. URL https://arxiv.org/abs/2410.19168.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, Ai Ti Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*, 2024. URL https://arxiv.org/abs/2406.16020.