AcousticBench

Measuring Acoustic Perception in Large Audio Language Models

Anirudh Rahul*, Arth Vidyarthi, Michael Horgan, Suyash Jaju*, Yuhao Zhou

David AI

CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 2026
EVALUATIONS AND DATASETS TRACK

CONTRIBUTIONS

A new benchmark of acoustic perception: 3,200 questions grounded in sound. Targeted fine-tuning narrows the human performance gap with the audio encoder frozen.

01 The benchmark

AcousticBench tests whether large audio-language models (LALMs) use pitch, loudness, vocal delivery, and recording quality to answer questions.

3,200 questions

5,355 unique recordings; 5,800 clip slots; 11.6 hours. Two choices per question, with A/B answers balanced within each full task. Chance: 50%.

RAD — 1,200 pairs

Relative Acoustic Discrimination

Compare pitch, loudness, tempo, or dynamic range.

Example: Which clip has higher pitch?

Speech, instruments, and synthetic signals; controlled acoustic differences.

SER-Classify — 600 clips

Speech Emotion Recognition

Identify an emotion from a single voice recording.

Example: Which of two emotion labels matches the voice?

Seven categories; English scripts chosen to limit lexical cues.

SER-Intensity — 400 pairs

Emotional intensity

Compare the strength of emotional delivery.

Example: Which clip expresses the emotion more strongly?

Same speaker and script, recorded at normal and strong intensity.

RSQ — 1,000 pairs

Relative Speech Quality

Compare real clean and degraded speech.

Example: Which recording sounds cleaner?

WebRTC conversations in 17 languages; pairs matched on duration and transcript length.

HUMAN INTER-RATER AGREEMENT

Annotators agree more on physical cues than on emotion categories.

Task	RAD	SER-Classify	SER-Intensity	RSQ
Krippendorff's α	0.96	0.69	0.74	0.82

Note. Nominal agreement on analyzed items; rating coverage varies. Values rounded to two decimals. Agreement is not an accuracy ceiling.

Labels grounded in audio

RAD labels are set programmatically. Independent annotators validate the emotion and quality items from audio alone. A separate listener pool supplies the human reference.

Examples above illustrate each question type. SER uses newly collected voice acting; RSQ preserves naturally occurring noise, codec and channel artifacts.

02 Model performance

Pairwise prompt example (RAD)



We evaluate 15 LALMs; open-weight systems lag further behind.

OVERALL AND TASK-FAMILY ACCURACY (%)

Gemini-3.1-Pro leads at 84.5% overall; the human reference reaches 93.5%.

Model	Overall	RAD	SER	RSQ
Human	**93.5**	**99.0**	**87.8**	**93.6**
Gemini-3.1-Pro	84.5	90.0	79.6	83.9
Gemini-2.5-Pro	77.4	74.1	74.2	83.8
Qwen3-Omni	69.4	70.7	69.5	68.0
Kimi-Audio	66.2	70.4	72.5	55.7
Qwen2-Audio-7B	51.3	49.9	54.1	50.0
MiMo-Audio-7B	54.1	50.5	61.7	50.0
ASR cascade	51.5	52.7	48.1	53.6

Note. Selected models; three-run means on full task sets. Overall averages RAD, SER, and RSQ. SER weights Classify and Intensity by question count; human scores pool ratings. Qwen2: Instruct.

ACOUSTIC CUE ACCURACY (%)

Gemini-3.1-Pro is stronger on pitch and tempo than on loudness and dynamic range.

RAD cue	Gemini-3.1-Pro	Human
Pitch	95.1	98.3
Tempo	98.2	99.5
Loudness	82.4	99.5
Dynamic range	86.7	99.0

Note. Full-set RAD accuracy; three-run model means and pooled human ratings.

EMOTION RECOGNITION AND INTENSITY (%)

Kimi-Audio is stronger on categories; Gemini and Qwen are stronger on intensity.

Model	Classify	Intensity
Human	87.1	91.4
Gemini-3.1-Pro	75.9	85.1
Qwen3-Omni	64.0	77.8
Kimi-Audio	74.0	70.2

Note. Full-set SER accuracy: 600 Classify questions and 400 Intensity questions.

Checking transcript shortcuts

The ASR cascade scores **51.5% overall**. Gemini-3.1-Pro's SER-Classify accuracy falls from **75.9% with audio** to **55.1% with transcripts**.

Cascade: GPT-4o-transcribe followed by GPT-5.4. Some lexical leakage remains; transcript-only results are diagnostic, not part of SER aggregates.

03 Adaptation and transfer

Freeze the audio encoder and base weights; train one decoder LoRA per task. All three models improve on all four in-domain tasks.



IN-DOMAIN ADAPTATION (%)

Decoder LoRA improves every task. Values show base → adapted accuracy.

Task	Kimi-Audio	Qwen3	MOSS	Human reference
RAD	69.2 → **98.7**	72.0 → **98.0**	66.5 → **90.8**	99.0
SER-Classify	70.3 → **75.2**	63.3 → **78.0**	60.7 → **74.7**	87.1
SER-Intensity	72.3 → **92.5**	79.0 → **94.0**	66.5 → **79.0**	91.4
RSQ	54.8 → **98.4**	69.4 → **99.4**	68.8 → **85.6**	93.6

Note. Kimi-Audio, Qwen3-Omni, MOSS-Audio. Model-specific recipes; 50/50 question-ID splits. Held-out n: RAD 600, Classify 300, Intensity 200, RSQ 500. Base and LoRA use the same held-out items; humans are the full-set reference. Splits can share speakers and source clips.

Speaker-disjoint retraining — Kimi-Audio

Gains persist on rebuilt splits: RAD **66.5** → **93.7**, SER-Intensity **70.5** → **90.0**, RSQ **61.5** → **95.5**. SER-Classify: 75.2 → 80.0; the gain is not significant.

Different train/evaluation items prevent attributing changes in gains to speaker overlap alone.

CROSS-TASK TRANSFER (%)

Compare RAD and lexical AudioQA training on new Kimi-Audio tasks.

Training	SER-Intensity	RSQ
Base Kimi-Audio	70.2	55.7
RAD	**77.9** (+7.7)	**61.3** (+5.6)
Lexical AudioQA	73.0 (+2.8)	57.0 (+1.3)

Note. Full target pools; no target-task training. Three-run means; gains in percentage points. Both adapters use the same A/B format, but training data differ. Neither improves SER-Classify. MOSS gains: +0.3 to +3.4 points.

CROSS-DATASET TRANSFER (%)

AcousticBench-trained Kimi-Audio adapters transfer to new recordings.

External dataset	Adapter	Base → LoRA
SonicBench	RAD	64.0 → **68.0**
VoiceBank-DEMAND	RSQ	62.0 → **74.2**
CREMA-D	Intensity	84.2 → **89.2**
RAVDESS	Classify	79.7 → **81.8**
LISTEN_full	Classify	37.2 → 36.0

Note. Task-matched adapters; no external-dataset fine-tuning. Evaluation formats and chance levels differ; CREMA-D uses intended-intensity labels. Transfer is uneven, with a decline on LISTEN_full. Accuracy differences alone do not establish significance.

*Work done while at David AI.

Dataset: huggingface.co/datasets/davidai-labs/AcousticBench Code: github.com/withdavid-ai/acousticbench-neurips2026