Episode Book

EPISODE #6272

Teaching Audio Models to Hear Beyond Words

PUBLISHED	25 October 2026 at 18:08 UTC
DURATION	23:24
LISTEN	myweirdprompts.com/episode/audio-models-sound-captioning-43417/
HOSTS	Corn, Daniel, Herman, Dorothy, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.22
SOURCES	16 cited by the research (listed at the end)
BOOK GENERATED	26 October 2026 at 06:38 UTC

SYNOPSIS

Standard speech recognition treats background sounds as noise to discard, losing vital context like breaking glass or engine trouble. This exploration covers how audio-language models like MOSS-Audio, MiDashengLM, and Audio Flamingo learn to perceive complete acoustic scenes. By replacing monotonic speech targets with rich captions and explicit time tokens, these architectures bridge encoders to language models to track environmental events, musical structure, and subtle acoustic textures rather than just producing transcripts.

TRANSCRIPT



Corn

I rather like the idea of an audio model that notices the kettle instead of faithfully transcribing you talking over it.



Daniel

That's what I'm getting at. Are there models whose main job is understanding the sound itself? What's happening in a recording, what kind of scene it is, even the mood, rather than just getting the words out?



Herman

There are, and the useful distinction is how they're trained. The leading models that describe sound also transcribe speech. You won't find much of a strict speech exclusion club. What you do find is training that treats music, machinery, voices and silence as things worth understanding together.



Corn

We've spent plenty of time on speech recognition. Daniel's taking us sideways into everything the transcript leaves out. I want the crash behind the sentence, especially if you then say the installation went smoothly.



Herman

A transcript can be completely correct while omitting the most important event in the recording. An audio caption might say someone speaks calmly while glass breaks nearby, then falls silent. That requires attention to events, timing and relationships. The words alone won't get you there.



Daniel

Does that distinction actually show up inside the models, or is it mostly a different prompt on the same system?



Herman

It shows up in the training targets and sometimes the encoder architecture. Two broad approaches are useful here. Audio-language models connect an audio encoder to a language model and generate descriptions or answers. Audio encoders without that language decoder produce embeddings, numerical representations you can use for retrieval, classification or another downstream task.



Corn

Although those encoders aren't necessarily speech-free. They can learn useful things about speech without turning it into a transcript. CLAP even learns relationships between sound and text. No talking decoder doesn't mean no language ever touched the training.



Herman

That's an important correction to the word pure. Dasheng, CLAP, PANNs and AudioMosaic belong in this conversation, but they aren't interchangeable and they aren't all trained the same way. Their common feature here is that you don't ask the bare encoder a question and receive a paragraph.



Daniel

What changes when the training target is a caption rather than a transcript?



Herman

MiDashengLM makes a deliberately provocative argument. It calls speech recognition pretraining a trivial objective because alignment is largely monotonic. Spoken words proceed in order, and the text follows them. A model can learn those local correspondences without learning much about the global acoustic scene. Trivial is their word. Recognising speech in difficult conditions is plainly hard.



Corn

The exam asks for words. The model sensibly ignores the forklift.



Herman

The loss function rewards that behaviour. A transcript usually leaves out reverberation, recording quality, music, environmental sounds and silence. It also omits vocal qualities such as apparent age, emotion or gender, although inferring those introduces its own uncertainties. General captions reward retaining more of that information.



Herman

Corn, would you count a good description of the sounds as understanding, or would you demand an explanation of what caused them?



Corn

I'd separate those. A short metallic impact followed by rolling is a defensible description. Someone dropped a saucepan is a hypothesis. I'd want a model that can stop before the saucepan, however much its language decoder wants a kitchen.



Daniel

Then which models should I actually associate with this category?



Herman

MOSS-Audio is a concrete starting point. OpenMOSS, MOSI.AI and Shanghai Innovation Institute released it in April. There are four-billion and eight-billion versions, each with Instruct and Thinking variants, under Apache two point zero. Environmental sound, music, captioning and time-aware questions are explicitly part of its remit.



Corn

Thinking sounds particularly useful when the sound is thinking.



Herman

It means the answering style, before you submit a recording of your afternoon nap. Also, model size labels aren't always complete checkpoint counts. The eight-billion Instruct checkpoint contains about nine point one billion parameters overall. The label is a family



Herman (continued)

designation, not a memory allocation promise.



Daniel

Is MiDashengLM the one that takes the caption approach most directly?



Herman

Xiaomi Research's seven-billion model is particularly explicit about it. It uses the open Dasheng encoder and general audio captions from ACAVCaps, built from public data. Instead of treating speech, music and environmental sound as unrelated annotation jobs, the captions combine them into one description. That changes what the system is rewarded for noticing.



Corn

A voice can be intelligible and nearly irrelevant. Someone says hello while an engine struggles to start. If your training target is hello, you've rewarded ignoring the engine's entire little tragedy.



Herman

The Audio Flamingo family from NVIDIA and the University of Maryland is another major branch. Audio Flamingo Two is a three-billion model. Audio Flamingo Three uses the unified AF-Whisper encoder and supports ten-minute audio. That gives it a longer scene to reason across, rather than an isolated event.



Herman

Audio Flamingo Next moves to eight billion and supports thirty-minute inputs, trained on over a million hours of curated audio. It also grounds reasoning steps to timestamps. Music Flamingo specialises further, with training covering harmony, structure, timbre, lyrics and cultural context.



Corn

You were going to choose Music Flamingo before the word harmony.



Herman

I was going to ask how it handles transitions. Different weakness.



Corn

You remain a DJ even during model selection.



Herman

Someone has to care whether the bass arrives before the explanation. Would you trust a music model that names all the instruments but misses the change in mood, Corn?



Corn

For cataloguing stems, possibly. For describing a whole piece, no. Naming a piano doesn't tell me whether it's carrying the melody or sitting behind something else. I'd care about its role changing over time. Otherwise the description is a receipt for instruments.



Daniel

Walk me through the machinery that gets from the recording to that kind of answer. Where does the language model enter?



Herman

MOSS-Audio gives us a useful representative architecture. Audio encoder, modality adapter, language model. Its encoder is trained from scratch, rather than simply borrowing an off-the-shelf speech frontend. It produces continuous temporal representations at twelve point five hertz. The adapter projects those into a form the language model can process, and the decoder generates text autoregressively.



Corn

Twelve point five representations a second sounds alarmingly coarse for a cymbal hit.



Herman

Those representations aren't twelve point five raw audio samples. They summarise much richer acoustic processing. The distinction matters. Still, compression can lose brief details, which is why the connections between encoder and decoder are interesting.



Daniel

Can the decoder recover details that the encoder has already thrown away?



Herman

No. It can guess plausible details, which is precisely what we don't want. MOSS uses DeepStack cross-layer injection to preserve access to different levels of acoustic features. Earlier and intermediate encoder layers get their own projections into early language-model layers, alongside the high-level representation.



Corn

Rather than only forwarding the encoder's final opinion.



Herman

Later layers can be good at broad meaning while losing prosody, transients and local time-frequency structure. Earlier features retain information useful for rhythm, timbre and background texture. A scrape and a tap might share a broad object category but differ in the detail that tells you what happened.



Daniel

Does that also explain how they answer questions about when something happened?



Herman

The time-marker design tackles that more directly. Explicit time tokens are inserted between audio frames at fixed intervals during pretraining. The model learns associations between the acoustic sequence and a textual time reference. That supports event localisation and timestamped speech recognition.



Corn

A model saying there was a bang somewhere in half an hour is describing my confidence in plumbing.



Herman

Timing changes the meaning too. A motor slows before the impact, or after it. The same inventory of sounds can support different interpretations depending on their order. Timestamp supervision gives the model a better opportunity to preserve that order. It doesn't guarantee causal reasoning.



Daniel

Where does the encoder-only approach fit? Is it mainly a cheaper version of this?



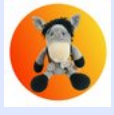
Corn

Well, before we get into that, we have a call from my mum.

D

Dorothy

Hello dear, it's your mother. It's Dorothy's Kitchen again, and today, sweetheart, we are making cold toast.



Herman

Hello, Dorothy. I'm listening.

D

Dorothy

You will need two slices of bread, and I would say a day old, no fresher, and butter, and then nothing else at all. Not a thing. That is the whole of it.

D

Dorothy

First you toast the bread. Watch it the entire time, dear. Do not look away, because the bread knows when you are not watching, and it will turn on you, it always does. Things have gone badly wrong here before and I would rather not go into how.



Corn

I'm here, Mum.

D

Dorothy

Then you butter it while it is hot, and then, and this is the solemn part, you wait. You leave it on the rack until every last bit of warmth has gone out of it. Timing is everything, and the waiting is longer than you think it will be.



Herman

On the rack throughout?

D

Dorothy

And that is cold toast. It is a disappointment, dear, but it is a reliable one, and sometimes that is what you want. Now eat something green today, and put a cardigan on, it is chilly out.



Herman

The encoder-only route can be cheaper, Daniel, but it's also a different output. AudioMosaic learns from patches of mel-spectrograms, using masked learning and contrastive learning. One objective learns to recover missing information. The other learns useful relationships between representations. It was pretrained on AudioSet, with no caption-generating decoder.



Corn

You can then attach a classifier or search by similarity. That might be all you need. A factory might want unusual bearing noise flagged against its normal recordings, not a paragraph about the bearing's emotional progression.



Herman

AudioMosaic transfers to environmental sound classification, speech commands and spoof detection tasks. Those are demonstrations of reusable representations, not evidence that one embedding solves every acoustic problem. MiDashengLM offers an efficiency claim on the language-model side too, up to four times faster time to first token and twenty times higher throughput than its comparison models.



Daniel

Those are quite different measures, though. Faster first answer doesn't necessarily mean keeping up with live audio?



Herman

Correct. Throughput, first-token latency, clip duration and streaming support are separate properties. Those are reported comparisons, not universal multipliers. A fast answer after uploading a complete clip doesn't tell you how quickly a system will detect an event while the recording is still happening.



Corn

The siren would prefer not to wait for the upload.



Herman

That has a concrete connection to accessibility. Captions for Deaf and hard-of-hearing users could describe traffic, approaching sirens and other environmental events. Industrial monitoring is another application. Both require recognising more than speech, and both care whether an answer arrives while it's still useful.



Daniel

Are those deployed uses, or applications people propose to justify a benchmark?



Herman

We should distinguish the two. SCENEBench grounds its evaluation in accessibility and industrial monitoring. That makes the tasks relevant, but it isn't proof of a deployed product's reliability. It contains roughly sixteen thousand clips across four task families, and it measures latency as well as task performance.



Corn

Accessibility also exposes a problem with polished captions. Siren heard is a different claim from ambulance approaching you. The second is more useful if correct, and more misleading if the model inferred direction from an emergency-scene cliché.



Herman

Spatial models address some of that information directly. MiDashengLM-Spatial, released this month, adds a dual-branch encoder for direction of arrival, distance and motion of overlapping sources. It points toward scene descriptions grounded in spatial audio. That still depends on the recording preserving the relevant cues.



Daniel

What else is useful beyond environmental alerts?



Corn

Searching recordings by what happens in them. Find the segment where machinery starts behind the speaker, or where applause interrupts the music. A transcript search can't retrieve an event it never wrote down. For a podcast archive, that could mean locating a sound without knowing which words surrounded it.



Herman

Music analysis is another good fit, especially instrumentation, structure and emotional progression. Audio questions over meetings can include overlap and delivery, not merely content. Robotics has uses for spatial scene understanding. Telehealth and well-being systems also target coughs or sobs in speech, though recognising a cough and establishing its medical cause are very different achievements.



Corn

Herman wants the clinical validation before the dashboard.



Herman

I want both. The dashboard tends to arrive first. Corn, would you accept an emotion label on one of our recordings?



Corn

I'd accept sounds frustrated more readily than is frustrated. Even then I'd want the audible basis, raised volume, clipped delivery, whatever it actually detected. A confident claim about someone's inner state can easily be a confident claim about the microphone or their accent.



Daniel

How well do the evaluations separate those kinds of understanding from just recognising words?



Herman

Often poorly. SCENEBench makes that criticism explicitly. Much evaluation measures what words were said rather than how they were said. It uses Qwen2-Audio's advertised ability to identify speech, music and ambient sounds as an example of the gap between a broad capability claim and what the usual tests establish.



Corn

The brochure hears the whole room. The exam asks for the sentence.



Herman

Some SCENEBench tasks fell below random chance across the five audio-language models evaluated. That doesn't mean every model is worse than guessing at every acoustic task. It does mean familiar leaderboard success can coexist with systematic mistakes on the properties those applications need.



Daniel

Below chance suggests something more than difficult audio. Could they be following misleading clues?



Herman

Potentially, yes. Answer preferences, linguistic priors or misunderstanding the task can produce systematic errors. I wouldn't assign one cause without inspecting the individual tasks. The result tells us to examine the mistakes, rather than average them into a reassuring overall score.



Corn

A model might be excellent at matching an answer to a familiar scene description and weak at identifying the acoustic evidence for it. Those abilities can look similar until the options change.



Herman

They do change. Multiple-choice scores shift substantially when answer options are reordered or paraphrased. The audio remains the same. Some of the measured performance therefore depends on how the question is packaged.



Daniel

Then is comparing model scores still worthwhile?



Herman

Within limits. MOSS-Audio's own comparison table gives its eight-billion Thinking model an average of about seventy-one across four audio benchmarks. Gemini three point one Pro is around eighty, while GPT four o Audio is around fifty-nine. That's a substantial separation under those test conditions.



Corn

The decimal places have better posture than the evidence.



Herman

They're useful measurements, just not measurements of everything people infer from them. Audio Flamingo Three and Next sit around fifty-eight and sixty on that table. Meanwhile, a four-billion MOSS model beats Qwen3-Omni and Step-Audio-R1 models roughly eight times its size on MMAU-Pro. Specialised audio training appears to matter a great deal.



Daniel

Is that enough to say caption training beats transcript training?



Corn

Not cleanly. The systems differ in architecture, data, optimisation and evaluation behaviour. It's evidence against buying parameter count as a substitute for acoustic training. Isolating the captions would require a more controlled comparison.



Herman

Mm. I'd narrowed that too far. Caption training is a strong explanation to investigate, not the only variable. There's a striking timing result too. On timestamped ASR, MOSS eight-billion Instruct has an alignment error score around thirty-six, against comparison



Herman (continued)

scores in the hundreds. Lower is better.



Corn

The time markers appear to be earning their keep.



Herman

The result is consistent with that design helping, although it doesn't isolate the component. Its speech-captioning score also edges ahead of Gemini three point one Pro in a language-model judge evaluation. That measures descriptive quality through another model's judgement, so fluency and listening still need disentangling.



Daniel

What happens when there is speech telling the model what it ought to hear?



Herman

That produces a remarkable weakness. Audio Hallucination Attacks inserts synthetic speech describing events that aren't actually present. Under the tested attack, Audio Flamingo Three was successfully misled about ninety-five percent of the time. Gemini three Pro was around eighty percent.



Corn

The dog has acquired a publicist instead of a bark.



Herman

Speech is itself legitimate audio evidence, so the model has to keep two claims separate. Someone says a dog is barking, and a dog is audibly barking. The attack exploits the collapse between those claims. Those percentages describe that adversarial setup, not ordinary error rates on everyday recordings.



Daniel

A caption could faithfully report the spoken claim while refusing to assert the event.



Corn

It could, and I'd find that more valuable than a beautifully complete scene invented from somebody's narration. The output has to retain where each claim came from. Otherwise adding words to a recording can overwrite the sound you're asking about.



Hilbert

Herman, would you call a sound rhythmic if the man making it was deliberately trying not to keep time? You've been careful about the order of events. I'm interested in the description on the label. My aunt Vera used to put irregular on everything.



Herman

Repetition can have a rhythm without being evenly spaced. Irregular alone leaves quite a lot unspecified.



Hilbert

She labelled effects at the little sound library where I worked in the mid eighties. I checked the descriptions against the tapes. We had a commission for twelve seconds of a municipal lift feeling uncertain. Vera hired a lift for the afternoon and read its inspection report to it through the doors. I recorded the mechanism. There was a low hum, two relay clicks, then nothing. She wrote hesitant ascent. It never moved.



Corn

The ascent was an ambitious annotation.



Hilbert

The client accepted it. They rejected my description because I had put the silence first. I wanted a portable recorder for that sort of job. Priced one at four hundred and sixty pounds. Didn't buy it. We took the library machine. Vera charged the client for the lift's



Hilbert (continued)

listening time, but not the reading. She said the report was public information.



Herman

Those two clicks would support a much narrower description than ascent. The silence between them might be important too.



Hilbert

That's the bit I like. Two tiny clicks can make you wait for a whole machine. You don't need a speech in there. Vera insisted the second click sounded less committed. I wrote second relay click, quieter than first. We argued over quieter for forty minutes because she'd turned the playback down halfway through. The final label was lift mechanism, no movement. The librarian had to find it again when another client asked.



Corn

... A searchable description has to survive being separated from the person who wrote it. An honest evaluation ought to test that, can somebody locate the event and distinguish what's audible from what's merely suggested? Below-chance tasks make me less interested in another grand average and more interested in whether those distinctions survive unfamiliar recordings.



Herman

Caption training and spatial encoders give us promising tools for that. The models Daniel is asking about are real, and scene description is a useful product category in its own right. Whether our tests reward careful listening remains an open question.



Corn

Our producer is Hilbert Flumingtop. Thank you for keeping the show running.



Herman

If you want more of this, try episode seven, Building Custom ASR Tools; episode three, Safetensors or something else; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts, the human-AI collaboration podcast. If you enjoyed listening, leave us a review wherever you get your podcasts.



Daniel

Thanks for having me in. See you soon.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **OpenMOSS, README, commit 66326e6 [primary]**
   https://github.com/OpenMOSS/MOSS-Audio

2. **model card, created 2026-04-13, modified 2026-09-25 [primary]**
   https://huggingface.co/OpenMOSS-Team/MOSS-Audio-8B-Instruct

3. **MOSS-Audio Technical Report, 2026-06-01**
   https://arxiv.org/abs/2606.01802

4. **MiDashengLM, v3 revised 2026-03-26**
   https://arxiv.org/abs/2508.03983v3

5. **MiDashengLM-Spatial, 2026-10-08**
   https://arxiv.org/abs/2610.11156v1

6. **Audio Flamingo Next, 2026-04-13**
   https://arxiv.org/abs/2604.10905v1

7. **Audio Flamingo 3, v2 2025-07-28**
   https://arxiv.org/abs/2507.08128v2

8. **Audio Flamingo 2, 2025-03-06**
   https://arxiv.org/abs/2503.03983v1

9. **Music Flamingo, 2025-11-13**
   https://arxiv.org/abs/2511.10289v1

10. **SCENEBench, EACL 2026, 2026-03-10**
https://arxiv.org/abs/2603.09853v1

11. **Audio Hallucination Attacks, 2026-03-31**
https://arxiv.org/abs/2603.29263v1

12. **Robustness of LALMs in MCQA, Interspeech 2026**
https://arxiv.org/abs/2510.04584v2

13. **ICML 2026 code repo**
https://github.com/HanxunH/AudioMosaic

14. **EvA: Evidence-First Audio Understanding, EMNLP 2026 Findings**
https://aclanthology.org/2026.findings-emnlp.2047/

15. **SCENEBench PDF (EACL 2026)**
https://aclanthology.org/2026.eacl-long.335.pdf

16. **DashengTokenizer**
https://www.alphaxiv.org/abs/2602.23765