Episode Book

EPISODE #6496

How Speech Models See Audio as an Image

PUBLISHED	8 November 2026 at 14:32 UTC
DURATION	30:01
LISTEN	myweirdprompts.com/episode/speech-recognition-log-mel-spectrogram-34138/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	9 November 2026 at 08:09 UTC

SYNOPSIS

Speech recognition models like Whisper never actually hear audio as human listeners do. Before a neural network processes anything, incoming waveforms are resampled, cut into twenty-five-millisecond windows, and compressed into log-mel spectrograms. This breakdown walks through every stage of that front-end pipeline, exploring the 1937 five-person study behind the mel scale, the deliberate discarding of phase and fine spectral detail, and why mismatches in sample rate or gain can derail transcription downstream.

TRANSCRIPT



Corn

The model never hears the audio. That's the sentence Daniel sent us, more or less — a speech recognition model never hears audio the way we do. Before the network sees anything, the waveform gets resampled, chopped into frames, and turned into a log-mel spectrogram, and models like Whisper then work in fixed thirty-second windows. He wants the full tour: what exactly happens in that front end, why the mel scale got chosen, what information survives and what gets thrown away, and how sample rate, microphone, gain and compression upstream quietly change what the model sees and how accurately it transcribes.



Herman

Herman Poppleberry here, and I've been waiting for this one, because the front end is where everyone stops reading the paper. They see "log-mel spectrogram" in the methods section and their eyes slide right past it to the attention diagram.



Corn

You've been waiting for a feature extractor. That's the bar you've set for excitement.



Herman

It's where the whole pipeline is decided! Everything downstream is downstream of that choice. Let's start with the assembly line, because once you walk it step by step, the philosophical questions answer themselves.



Corn

Go.



Herman

Step one, resampling. Whisper takes all incoming audio and resamples it to sixteen kilohertz, mono. And that's the first irreversible act, before the network exists, before anything smart happens. Sixteen kilohertz sampling means a Nyquist limit of eight kilohertz. Everything above eight kilohertz is gone. Permanently. Sibilance that lives up around ten, twelve kilohertz, room character, the difference between a good microphone and a great one — discarded, because the bet is that speech intelligibility lives below



Herman (continued)

eight kilohertz.



Corn

Which is true for phonemes.



Herman

For phonemes, yes. Less true for the cues that tell you who's speaking, or what room they're in, or whether the recording is good. That's a decision about speech made on behalf of everything else.



Corn

The bouncer at the door only lets in people below eight kilohertz.



Herman

Step two, framing. The waveform gets cut into twenty-five millisecond windows with a ten millisecond hop. In samples at sixteen kilohertz, that's a four hundred sample window, hopping a hundred and sixty samples each step. Each window gets a Hann window applied and a short-time Fourier transform. And twenty-five milliseconds is this beautiful classic compromise — a low voice has pitch harmonics spaced roughly forty hertz apart, so you need enough window length to resolve adjacent harmonics. Too short and the harmonics blur together into one lump.



Corn

And too long?



Herman

Too long and the vocal tract has moved during the window — you smear a consonant across a phoneme boundary, you average a "ba" and a "da" into mush. Twenty-five milliseconds is long enough to see pitch, short enough that the vocal tract is roughly standing still.



Corn

So every ten milliseconds, we take a photograph of the spectrum.



Herman

Then step three is the compression. Each of those magnitude spectra has hundreds of FFT bins, and it gets projected onto eighty triangular filters spaced on the mel scale. Hundreds of numbers become eighty. Then you take the log of the energies in each of those eighty bands. Then the whole thing gets globally normalized — scaled to roughly between minus one and one, roughly zero mean. And then the final, strange step: the audio is clipped or padded with silence to exactly thirty seconds. Four hundred and eighty thousand samples, which becomes three thousand frames by eighty mel bins. That three thousand by eighty tensor is the image. That is literally what the transformer encoder looks at.



Corn

It's a picture of the last thirty seconds.



Herman

A grayscale picture, eighty pixels tall, three thousand wide. And the encoder then runs a stride-two convolution over it that halves it to fifteen hundred positions, so at native resolution the model emits one token per twenty milliseconds of audio.



Corn

One word per...



Herman

Per twenty milliseconds of audio at the encoder's finest granularity — the decoder can merge across them, obviously, but that's the resolution of the internal grid.



Corn

Herman, I want to slow down on the log, because you said it does two jobs and then walked past both.



Herman

Right, no, that's the interesting bit. Job one: the ear's loudness perception is roughly logarithmic — a doubling of physical intensity is perceived as a modest increment, not a doubling. Taking the log of the band energies puts the feature scale in the same territory as perception. Job two is sneakier and arguably more important: log turns multiplicative changes into additive ones. If the recording gain doubles, the raw band energies all double. After the log, every band just shifts up by the same constant. The whole spectrogram slides up by a uniform offset.



Corn

Which is removable.



Herman

Cepstral mean normalization — subtract the average spectrum across time, and the gain offset cancels. That trick only works because of the log. Without it, a quiet recording and a loud recording would have different feature shapes.



Corn

I'd assumed normalization was a courtesy. It's the reason a whisper and a shout look the same to the model.



Herman

Within reason. We'll get to where it stops working.



Corn

Before we do — the mel scale. Eighty filters spaced on it. You said the word "mel" three times and haven't defined it once.



Herman

Because it deserves its own segment. The mel scale is a perceptual pitch scale — pitches judged by human listeners to be equal subjective distances apart. The reference point is that a thousand hertz tone, forty decibels above hearing threshold, is defined as a thousand mels. Below roughly five hundred to seven hundred hertz, the scale is close to linear — hertz and mels track each other. Above that, increasingly large hertz intervals produce equal mel increments. The ear compresses. Roughly logarithmic up top.



Corn

And who decided where it bends?



Herman

Ah. Stevens, Volkmann and Newman, nineteen thirty-seven, at Swarthmore. "A Scale for the Measurement of the Psychological Magnitude Pitch," Journal of the Acoustical Society of America, volume eight, page two hundred eight. The word "mel" is short for melody. And the method — this is the part I want you to sit down for...



Corn

I'm hanging by three limbs. Continue.



Herman

Half-value pitch determinations. Listeners adjusted a tone until it sounded halfway between two reference pitches. And the number of observers in that foundational study was five. At sixty decibels. Five people, in nineteen thirty-seven, twirling a dial.



Corn

Five.



Herman

Five. And the formula everybody quotes — mel equals two five nine five times log of one plus f over seven hundred — that's not even from the original paper. That curve fit is O'Shaughnessy, nineteen eighty-seven, and the seven hundred hertz corner constant traces back to Makhoul and Cosell in nineteen seventy-six. So the formula on the whiteboard is a nineteen eighty-seven retrofit of a nineteen thirty-seven experiment run on five people.



Corn

The entire field stands on the pitch judgment of five Swarthmore undergraduates.



Herman

Presumably undergraduates. And it's worse — or better, depending on your taste — the mel scale has about three studies total behind it, only two of which used half-pitch judgments. Compare that to the sone scale for loudness, which has a whole literature. Mel is one of the least-replicated results that became one of the most load-bearing defaults in engineering.



Corn

Herman.



Herman

Yes?



Corn

I'm going to say something nice about it, and I need you to not make it weird.



Herman

I'll contain myself.



Corn

Nobody has re-run the experiment because nobody needed to. Every speech system built since uses mel or something like it, and they work. The five observers were right, or right enough that eighty years of engineering never had a reason to fire them.



Herman

That's the fair counterweight, and there's a physiological argument underneath it too — the real justification for mel-ish spacing is critical bands. The cochlea's basilar membrane does roughly logarithmic place coding of frequency. Mel-spaced filters approximate the ear's own channelization — you're giving the model roughly the frequency resolution a cochlea has. So the general shape is grounded. But the specific implementation — eighty bins, the seven hundred hertz corner, those exact triangle widths — that's convention. It's physiology's rough outline, filled in by thirty years of MFCC pipelines and inertia.



Corn

A police sketch of the ear, drawn from a description.



Herman

Drawn from five witnesses, in one room, in the thirties. And it's mostly on file because nobody's wanted to pay for a new sketch.



Corn

Okay. So that's the scale and where it came from. Now the other half of Daniel's question — what does the model actually keep, and what does it throw away?



Herman

The surviving list first. Short-time spectral magnitude envelope — the formant structure, which carries phoneme identity. That's the core payload. Coarse temporal dynamics at ten millisecond resolution, which is plenty, because phonemes run fifty to two hundred milliseconds. Relative energy across eighty bands. There is enough in that tensor to reconstruct intelligible speech in principle — people have run vocoder experiments essentially inverting these features.



Corn

And the discard list is longer.



Herman

Much longer. Absolute phase, first. The mel spectrogram is magnitude-only — phase is thrown out entirely. Every frame is a photograph of how loud, not of which way.



Corn

Does that matter? I've always half-assumed phase was decoration.



Herman

For clean single-channel speech, less than you'd think. But it's not nothing — one finding from that Amazon Opus work was that codec-preserved phase still helped a learned beamformer. So phase carries real information about direction and source separation; it's just not in the features. Then, fine spectral detail within each mel band — everything inside a filter is averaged. This is exactly why mel features are poor for music and for speaker identity. Then everything above eight kilohertz, gone at the resample. Absolute amplitude, mostly absorbed by the log and normalization. Precise timing below ten milliseconds — onset transients, fine voice onset time detail. And long-range structure — the thirty second window is a hard horizon. Nothing crosses it.



Corn

The model is a very good reader who's been handed a summary.



Herman

A summary written to a template, too. Every loss in that chain is justified by what human hearing supposedly ignores — we don't miss phase much, we don't resolve fine spectral detail within a critical band, loudness is logarithmic. The front end is a chain of deliberate losses, each one excused by the ear. The question Daniel's really poking at is whether the ear's excuses are good ones, and whether the excuses hold for the things we now ask these models to do.



Corn

And that gets us to the upstream half, because everything you've described assumes the audio arriving at the front end is clean. Daniel's question is what happens when it isn't — sample rate, microphone, gain, compression.



Herman

And this is where there are actual numbers, which I've been saving. Sample rate first. A wideband model trained at sixteen kilohertz degrades substantially on eight kilohertz telephony audio, and the failure is attributed directly to missing information above four kilohertz. So you get a double penalty — the information is gone, which is physics, and then the model is mismatched, because it was trained expecting those bands to contain something. It's not just deaf up there, it's confidently wrong about what it expected to hear.



Corn

Eight kilohertz telephony audio — that's plain old telephone.



Herman

Narrowband telephony, yes. And the codec question extends it, because nobody ships raw audio anymore. There's an Amazon study — Khare, Sundaram and Wu — that compressed audio with Opus at different bitrates and measured relative word error rate degradation against uncompressed. And the table is a cliff. At one hundred twenty-eight kilobits per second, plus two point four percent — basically nothing, inaudible. At thirty-two, plus six point six. At sixteen, plus twelve point six.



Corn

And at eight?



Herman

Plus two hundred two percent.



Corn

Two hundred.



Herman

Two hundred two point one. A tripling of the error rate. And the reason is precise — Opus drops to narrowband encoding at eight kilobits. No information above four kilohertz at all. Same cutoff as the telephone. The cliff isn't a gradual codec-quality thing, it's a second Nyquist event happening inside the codec.



Corn

So the front end throws away everything above eight kilohertz, and the codec, if you're stingy, throws away everything above four — and the model trained on the first world can't function in the second.



Herman

That's the shape of it. One caveat worth saying: those numbers are from a far-field multi-channel system, not Whisper — the direction and rough magnitude transfer, the exact percentages don't.



Corn

So the degradation is real and quantified. Can you train your way out of it?



Herman

Partially — and this was my favorite result in the whole pile. They trained on mixed bitrates, so the model saw compressed and uncompressed audio during training. That recovered up to half the degradation. At thirty-two kilobits, the six point six percent dropped to three point three with a model trained on that exact bitrate, and four point one with one general mixed-bitrate model.



Corn

Robustness for the price of one training run.



Herman

But there's a bill. The mixed-bitrate model cost three point four percent error on clean uncompressed audio. You bought robustness with clean-audio accuracy — the model spends capacity learning to tolerate distortion it wouldn't otherwise need to model. Although, in fairness, at sixteen kilobits the mixed model actually beat the matched one, seven point nine against ten, because the matched model never heard the higher-frequency information that even sixteen-kilobit Opus preserves — training matched to a degraded input can leave you knowing the degradation but not the speech.



Corn

You become an expert in the smear and a tourist in the voice.



Herman

Gain is next, and it's subtler. The log plus global normalization absorbs a lot — a quiet recording and a loud one land in similar feature ranges, which is why you can whisper at a dictation app and it mostly still works. But it's not unlimited. Hard clipping is the hard case: when the signal hits the ceiling and flattens, you generate broadband harmonic distortion — spurious energy that lands inside the mel bands. The model has no way to distinguish that from real fricatives.



Corn

Because fricatives are broadband noise too.



Herman

Noise is exactly what a clipped peak looks like in the spectrum. And aggressive dynamic-range compression — the kind of leveling that cheap recorders and every conference-call system applies — flattens the amplitude envelope, and low-energy consonants get smeared down into the noise floor. Anything nonlinear upstream creates artifacts, and the front end is linear-plus-log. It can normalize a quiet room. It cannot un-lie a clipped one.



Corn

I once recorded a whole episode segment into a mic with the gain pinned red, and the transcript came out describing a kitchen. There was no kitchen.



Herman

I remember that transcript. You were allegedly making soup.



Corn

Allegedly. The microphone had opinions. Which is the last item — the mic itself.



Herman

Frequency response is the obvious spec, but it's honestly the least of it. The two that dominate are the self-noise floor and the placement. A cheap microphone's hiss sits in exactly the same mel bands as fricatives — that constant electronic noise floor is spectrally shaped like the consonants the model needs. And directivity and distance change the direct-to-reverberant ratio. Get far from the mic in a hard room and the reflections arrive inside your twenty-five millisecond frames, smearing them. The Amazon paper's whole motivation was the multi-device world — a person whose recordings come from different manufacturers, different encodings, different bandwidths, at different times. Which is now everybody.



Corn

Every phone is a different microphone for the same voice.



Herman

And the model has to be all of their models at once.



Corn

There's one piece of Daniel's framing we haven't touched — he called the thirty-second window "fixed," and there's a story hiding in that word.



Herman

There is, and it's the best story in the whole topic. The encoder doesn't accept variable-length input. Every clip — two seconds, half a second, whatever — is zero-padded with silence to exactly thirty seconds. A two-second utterance is two seconds of speech plus twenty-eight seconds of digital silence, presented as one image. Batching convenience. Nobody chose it for acoustic reasons.



Corn

So silence is a first-class input.



Herman

Silence is a fully legitimate, expected, everyday input. And Whisper hallucinates on it. There's a test set of one thousand eight hundred ten clips containing no speech at all — forty-two room tone, six hundred music, about eleven hundred label-verified non-speech sounds. Every OpenAI checkpoint produced output on every silence clip. Not most. Every one. Large-v2, one hundred percent. Large-v3, one hundred percent. Turbo, one hundred percent.



Corn

Output of what?



Herman

"Thank you." "Thanks for watching." Subtitle boilerplate — because the model was trained on YouTube captions, and the most common text near silence in that corpus is an endcard. And the correct output for a no-speech clip is the empty string, so any output at all is wrong by definition.



Corn

A model trained on the internet, shown pure silence, thanks you for watching.



Herman

It's the most honest thing any of them has ever said. And the mechanism is the padding — the architecture guarantees silence is always in the picture, and the model has no clean concept of "this region is padding," so it treats drawn-in silence the same as recorded silence. When the recording really is silence, the model is in its most out-of-distribution territory, and it reaches for its priors, and its priors are YouTube endcards.



Corn

What's the fix people actually use?



Herman

External voice activity detection — gate the audio before it ever reaches Whisper, so silence never gets in. And there's a built-in heuristic, hallucination-silence threshold, added by a pull request in late twenty twenty-three, that skips decoding over quiet stretches. Which mostly works. Mostly. There was a bug report this year, one of the MLX ports, where the silence threshold, the no-speech threshold, and the log-probability threshold were all inert — every one of them — because the no-speech probability came back as not-a-number. A single NaN unwired the entire safety apparatus.



Corn

The guardrails were installed. The guardrails were not connected.



Herman

They were connected to a NaN.



Corn

Before I lose the thread entirely — and now it's time for Dorothy's History of Toast.

D

Dorothy

Hello dear, it's your mother. Dorothy's History of Toast.

D

Dorothy

In eighteen sixty-one, Mrs Beeton's Book of Household Management has a recipe for a toast sandwich. One slice of toast between two slices of bread, recommended for invalids.

D

Dorothy

I suppose that is the trouble with toast. You go to all that bother, and then it is a sandwich again.

D

Dorothy

Anyway. Your father has left his teacup on the windowsill again. Be good, sweetheart.



Herman

Mum, the toast sandwich is real, I've looked it up before — forty calories of pure Victorian anxiety.



Corn

She's not wrong about the bother, though. Thirty seconds of padding and you've got a sandwich again.



Herman

Let's get back to the codec cliff, because there's a second-order point I don't want to lose. The Opus numbers are a compression penalty measured on clean recordings. But in the real world, the bitrate decision correlates with the device — the cheap headset runs the cheap codec — so the microphone noise, the narrowband cutoff, and the codec artifacts all arrive stacked. Daniel's question lists those as separate factors, and upstream they never are.



Corn

They're one purchase.



Herman

They're one purchase at one price point. The failure isn't additive either — the noise floor is inside the bands that survive, the codec distortion is inside the bands that survive, and the model was trained on recordings where none of that was there. Every layer assumes the layer above it behaved.



Corn

And the front end can't audit the layers. It just reports eighty numbers per frame and lets the network guess what world produced them.



Herman

Which is the answer to "how do upstream factors change what the model sees" in one line — they change it before the model can even express an opinion about it. There's no error signal for "this was clipped at the preamp." The spectrogram of clipped speech is a legal spectrogram.



Corn

It's the same picture a different truth would have produced.



Herman

That's the trap. A corrupt input that looks like a legal input is the worst case, because nothing downstream can flag it. And the fixes all push the problem around rather than removing it — you can train on mixed bitrates, and pay three point four percent on clean audio. You can VAD-gate the silence, and depend on the gate. You can spec better microphones, and the users are the microphones, and the users are the cheapest ones.



Corn

Which is why the two-second clip of me and the soup ended up as a transcription of a kitchen. Not a model failure — a chain of small decisions, each reasonable, compounding into fiction.



Herman

And the fiction is fluent, because the model's only escape from a corrupted input is through its priors, and its priors are the internet.



Hilbert

Which of you said five observers?



Corn

Both of us, I think. Why?



Hilbert

Herman, then. You kept saying the ear is logarithmic above five hundred hertz, like the ear filled out a form. I've done that experiment. Not the Swarthmore one. Mine. A signal generator, a set of tuning forks, a spectrum analyzer a man at the allotment club lent me, and a Tuesday. The room sounded wrong. That was the whole dispute. A chap called Reginald insisted it was the radiator. I said it was a resonance at two hundred hertz and I



Hilbert (continued)

was going to prove it with numbers.



Herman

Was it the radiator?



Hilbert

It was a radiator with a resonance, which is the worst kind, because you're both right and nobody apologizes. But the mel business, that's what I came out about. Five observers is generous. You try adjusting a tone until it sits halfway between two others, Corn, and after the first twenty minutes your judgment drifts like a wheel with a bent rim. I logged it. Halfway through my Tuesday, my halfway points moved. Me, the same ears, an hour apart. The notebook's got the numbers.



Corn

The notebook exists?



Hilbert

It's in the shed. Nobody's read it — not private, it's just that when I put it down the pages were in the wrong order, and I'm not having people reading a document I know is shuffled. I'll re-sort it first. That was three years ago. The results were significant, though. Statistically significant, personally, which is the only sample that reports to me.



Corn

One man, one notebook, self-significant.



Hilbert

It's more than five if you count each Tuesday. I did four Tuesdays. Reginald calls the analyzer "the listening machine," which is why I stopped lending him things.



Herman

The drifting halfway points — that's actually consistent with what's known about the psychophysics. Pitch-matching judgment fatigues. Those nineteen thirty-seven observers were presumably fresh, and five fresh observers is a sharper instrument than me at hour two.



Hilbert

You'd want them fresh, yes. The funny thing was the tuning forks were the reliable end and they weren't even in the disputed range. Held one to the grille with a claw tip, which you're not supposed to do, and it still told you more than the analyzer's own trace, which wobbled whenever the fridge cut in.



Corn

The fridge was on your channel.



Hilbert

The fridge was on every channel, Corn. Two hundred hertz, give or take, same as the radiator, which is how the argument started, if you're keeping up. Reginald got the analyzer back with a dent I don't discuss.



Herman

Corn, that line you had earlier — the model was trained on the internet, shown silence, and thanked you for watching. I keep coming back to what Hilbert's fridge and radiator actually did to his measurements, because it's the same structure. A constant contaminant inside the band you're measuring, and no way to subtract it after the fact.



Corn

Except Hilbert could go stand next to the fridge and notice. The model gets one picture.



Herman

One picture, no Tuesdays. It can't re-measure.



Corn

Here's where I've come out on Daniel's question, then. The front end isn't a passive window — it's a series of editing decisions made before the model has any say. Kill everything above eight kilohertz. Average what's left into eighty bands. Drop the phase. Normalize the loudness. Draw in silence to make the frame fit. Each decision has an excuse, and the excuses are mostly the ear's — but the ear got to vote with five observers in nineteen thirty-seven, and every device since has been living with the result.



Herman

And Daniel's upstream question is the other half of the same thing — the front end can't defend against what happens before it. The mel bands don't know they're being fed a clipped signal. The normalization doesn't know the noise floor is a microphone. The padding doesn't know the silence is real.



Corn

It's turtles, and the bottom turtle is five people in Swarthmore.



Herman

The part I find open is the exit. Models are already moving toward learned front ends — raw waveform in, let the network discover its own filterbank. And the interesting question is whether the thing they discover is the mel scale again, because the cochlea is real and logarithmic compression of frequency is good engineering, or whether they find something better and we find out the whole eight-decade default was a local maximum.



Corn

And if they do find something better, they inherit the same upstream fragilities — the codecs and the microphones don't stop being lossy just because the front end got smarter. Better glasses don't fix the fog.



Herman

Or, darker thought — they inherit new ones, because a learned front end is tuned to whatever it saw in training, and the training corpus is the internet, which leads us back to "Thanks for watching."



Corn

If the mel scale is a nineteen thirty-seven convention and the thirty-second window is a batching convenience, the question worth carrying out of this episode is what else in the modern stack is load-bearing purely by accident. Something is. Something nobody's looked at since the five-observer era.



Herman

Producer credit where it's due — Hilbert Flumingtop at the desk, as ever, dents notwithstanding. If you want more of this, try episode ten, How ASR Went From Frustration To ... Whisper Magic; episode eleven, How Does Fine Tuning Work Anyway; and episode three, Safetensors or something else. This has been My Weird Prompts. If the episode earned its thirty seconds of padding, leave us a review on Spotify — it's the silence where the feedback goes.



Corn

We'll be back soon. Mind the radiator.



Herman

And the fridge.