Episode Book

EPISODE #6055

How to Fine-Tune ASR on Noisy Audio

PUBLISHED	17 October 2026 at 13:23 UTC
DURATION	42:22
LISTEN	myweirdprompts.com/episode/asr-fine-tuning-noisy-audio-00244/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	upstage/solar-pro4
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	19 October 2026 at 17:43 UTC

SYNOPSIS

Building a robust speech recognition model feels like it should require collecting mountains of noisy audio. It doesn't. The research points to a different approach: build a compact noise library, mix it into clean speech at sampled signal-to-noise ratios, and keep a healthy share of clean originals in every epoch. This episode walks through the composition questions that actually determine success — how much augmented data is too much, which noise types to prioritize, why the best adaptation technique depends on whether the noise is stationary or not, and how to schedule augmentation across training. Along the way: the ROAR paper's original-to-augmented ratio, NVIDIA's Riva recipe, the IIT Bombay noise taxonomy, and a field-recorded noise dataset from 165 Indian districts that captures overlapping sound the way deployment actually delivers it.

TRANSCRIPT



Corn

Here is the thing about noise in automatic speech recognition. Most people think the trick is collecting more of it.



Herman

They think if they just record a hundred hours of a crowded street the model will wake up clearer.



Corn

Which is how you end up with a model that performs worse than when you started, because it forgot everything else.



Herman

Daniel is circling exactly that. He is building his own fine-tune and keeps coming back to background noise and speaker conditions. Can clean samples sit next to noisy ones, what is the right balance, how many noisy samples would you actually collect, which noise types would you go out of your way to capture, and is background conversation a confuser or the very signal that teaches the model to tell the intended speaker from the audio it picked up by accident.



Corn

The last one is the part that wavers. He can see it going either way.



Herman

He is right to feel that way, because the literature is unusually clear on the first four and split on the fifth. Which is why we can spend the whole episode on composition rather than just collection.



Corn

Daniel's question is a composition question wearing a collection question's coat.



Herman

Let's strip the coat.



Corn

You just put a coat on the question.



Herman

There is a reason.



Corn

Fine.



Herman

The first thing to say is that dataset composition for noisy ASR is not about collecting noisy speech one sample at a time. It is about building a compact noise library and mixing it into clean speech at sampled signal-to-noise ratios.



Corn

That reframe gets more mileage than most of the tools people buy.



Herman

It does because the noise is cheap. The two-shot that made this obvious was the IIT Bombay robust-ASR study. They built a usable noise set of about two hours total across seven noise types, ten training samples and eight test samples per type, and still got large word error rate gains.



Corn

Two hours to make a model noticeably less fragile. That is the kind of return on effort that makes you suspicious.



Herman

You should be suspicious, because it points to where the effort actually belongs. It belongs in diversity of noise types and signal-to-noise ratios, not volume of noise recordings. That is the spine of the answer before we even touch the babble question.



Corn

Which is good, because Daniel's last question is the one that would eat an afternoon if we started there.



Herman

The question is whether clean and noisy can coexist in the same fine-tune. The answer is yes, and it is standard practice rather than a compromise.



Corn

I would have guessed that was a compromise you make when you don't have enough clean data.



Herman

That was the old intuition. NVIDIA's Riva fine-tuning tutorial makes the current shape plain. They take a clean dataset, the AN4 corpus, and augment it with background noise at zero, five, ten and fifteen decibels signal-to-noise ratio, then train on the union. The clean originals stay in the set.



Corn

So they are deliberately keeping a foot in the clean room while standing in the noisy one.



Herman

And the reason they keep the clean data is the single most important mechanism in this whole topic.



Corn

Do tell me the mechanism before I assume it is a heuristic.



Herman

Catastrophic forgetting. Fine-tune exclusively on narrowband or noisy audio and the model gets better on that domain while quietly getting worse at everything else. Spirelight's telephony guide frames it this way: fine-tune only on the phone calls and the model gets better at calls while quietly getting worse at everything else, and that regression surfaces as a mystery bug weeks later.



Corn

A mystery bug that is just your model remembering less than it used to. That is a lovely thing to explain to a stakeholder.



Herman

It is also the kind of regression nobody attributes to the fine-tune, because the fine-tune was supposed to help. The lesson is that you blend a slice of clean wideband speech from the original training distribution into every epoch.



Corn

So the clean originals are not dead weight. They are the thing stopping the model from quietly amputating its knowledge of rooms it has not been in yet.



Herman

That is a slightly gruesome way to put it.



Corn

It landed.



Herman

The balance question then has a concrete answer, which is where the ROAR paper earns its place. It is by Singh and colleagues at the University of Eastern Finland, on the arXiv, and it defines the original-to-augmented ratio, the OAR.



Corn

A ratio with a name. Always prefer a ratio with a name.



Herman

The finding is that exceeding a certain amount of augmented data from certain methods beyond a certain limit degrades ASR performance, while keeping the augmented ratio too small does nothing useful. The best fixed ratio was OAR of one to two for small fine-tuning sets, and OAR of one for ten-hour and one-hundred-hour sets.



Corn

OAR of one is equal parts, one to two is original-heavy.



Herman

Right. And the counterintuitive part is their scaling finding: a lesser amount of augmented data is required as the amount of training data increases. More clean speech means you need less noise augmentation, not more.



Corn

That is the opposite of the folk wisdom.



Herman

It is, and it matters practically. If you follow OAR of about one, that implies roughly as many noisy hours as clean hours in the fine-tuning mix. But the noise itself remains cheap, because you are mixing from a small library.



Corn

So Daniel's question about how many noisy samples to collect dissolves a bit.



Herman

It dissolves into ratios and probabilities and hours rather than a count. No paper gives an absolute number of noisy samples to collect.



Corn

I was hoping for a clean integer. I would have framed it on a slide.



Herman

The practical anchors are the bands SpeechData.ai and Spirelight publish. Five to fifty hours of in-domain audio to fine-tune Whisper on a new domain, fifty to five hundred for accent and dialect work, and ten to twenty-five hours of matched call audio as a starting point for a telephony fine-tune.



Corn

Ten to twenty-five hours of matched call audio as a starting point, and Whisper itself was pretrained on six hundred eighty thousand hours for context. The scale mismatch between what made Whisper and what you fine-tune with is itself a lesson.



Herman

It is, and the ROAR dynamic schedule is the part people skip and then reinvent badly. They keep augmentation low at the start of training for better initialization, raise it mid-training to avoid overfitting, and taper it back to zero at the end.



Corn

You taper the noise away at the end of training.



Herman

Right.



Corn

So the model spends its last stretch being beautiful and quiet.



Herman

That is the correct move, and it is copyable directly. The ROAR dynamic schedule gave about a five percent relative word error rate improvement over an open-source wav2vec2.0 base.



Corn

Five percent on a base model with this much headroom is not nothing.



Herman

It is not. And the IIT Bombay study used an even simpler rule of thumb along the same lines. They augmented clean speech with noise with a probability of zero point five.



Corn

About half the training utterances get noise added.



Herman

That is a rule you can carry into a first build without having to think about schedules yet. Half the utterances get noise, half stay clean, and the clean originals remain.



Corn

So the first three of Daniel's questions already have answers that are not "it depends."



Herman

They do. Mixing clean and noisy is standard, the balance is a ratio rather than a count, and the number of noisy samples you collect is tiny relative to the speech because you build a noise library and mix from it.



Corn

I would have collected noise recordings until my hard drive groaned. That would have been the wrong effort in both directions.



Herman

The wrong effort is also the one most tutorials encourage, because they talk about collecting noisy speech samples and never about building a noise library. NVIDIA's tutorial uses thirty-second isotropic noise clips, the first fifteen seconds for training, the last fifteen for evaluation.



Corn

Fifteen seconds for training. If I met someone reducing a noise problem to fifteen-second clips I would want to see their room.



Herman

Their room is probably fine. The point is that a noise sample is a resource you reuse, not a one-shot recording of someone speaking in a cafe.



Corn

That distinction is the difference between a project that finishes and a project that becomes an audio archaeology expedition.



Herman

It is also the distinction between collecting samples and building a library. And the library is where the noise-type question lives, which is the next of Daniel's questions.



Corn

Which noise types would you go out of your way to capture. I am hoping for a ranked list I can pretend I planned all along.



Herman

The IIT Bombay taxonomy is the most explicit one, and it is a good starting shape. Babble, airport or station, car, metro, cafe, traffic, and air conditioning or vacuum. They split these into stationary and non-stationary.



Corn

Stationary being the ones that sit still.



Herman

Car, metro and traffic are the stationary ones. Babble, airport or station, cafe, and air conditioning or vacuum are the non-stationary ones.



Corn

The non-stationary ones being the ones that arrive dressed as if they have plans.



Herman

The split matters because the best adaptation technique differs by type.



Corn

Which is the detail I would have missed by making a single playlist of everything noisy.



Herman

That is the mistake. For the stationary noises — car, metro, traffic — speech enhancement with something like Demucs wins. For the non-stationary ones — babble, cafe, airport — adversarial training wins.



Corn

So "which noise to capture" and "how to handle it" are coupled decisions. You do not pick noise types in a vacuum, you pick them with your adaptation technique.



Herman

That is one of the better insights in the paper. Knowledge of the underlying noise type can meaningfully inform the choice of adaptation technique.



Corn

The sentence is almost polite enough to be from a paper. Almost.



Herman

There is a newer taxonomy that matters for people working in the field, the VAANI Noise Event Dataset, on the arXiv from September this year.



Corn

You do not sound like you have been waiting to say that.



Herman

I have not been waiting. It just appeared and it is useful. It offers a seven-class taxonomy from real field recordings across one hundred sixty-five Indian districts. Animal, traffic, baby or child, music, signal or alarm, appliance, and non-speech human.



Corn

Non-speech human as its own class is the one that makes me think of Daniel's last question again.



Herman

It is the right instinct. VAANI's notable feature is that it captures in-situ overlapping noise with timestamped spans, which is closer to how noise actually arrives in a deployment than a clean isolated clip of each type.



Corn

Timestamped overlapping noise as a dataset feature is the difference between an exercise and a record of what actually happened.



Herman

If you want standard noise corpora to draw from instead of building your own from scratch, there is a reliable roster. MUSAN, DEMAND, the OpenSLR RIRS_NOISES set, AudioCaps, FSD50K, AudioSet, WHAM!, and CHiME-6.



Corn

That is a list long enough to feel like a research universe and short enough to be real.



Herman

It is real, and if you are building a personal fine-tune the practical move is to take a couple of those, add what you can capture locally, and concentrate on the right mix of types and signal-to-noise ratios.



Corn

Signal-to-noise ratios being the thing NVIDIA samples at zero, five, ten and fifteen decibels, and the IIT Bombay study samples from zero through twenty-five decibels.



Herman

And ROAR samples zero to twenty decibels uniformly.



Corn

Everyone samples the range, which is reassuring in the way a crowded room is reassuring when you are trying to find a quiet one.



Herman

The reason the range matters is that the same noise type at different signal-to-noise ratios behaves differently. A cafe at twenty decibels is background texture. A cafe at zero decibels is the conversation you were trying to record.



Corn

That is the moment the babble question stops being theoretical. Daniel's last question is whether background conversation confuses the model or teaches it to separate speakers.



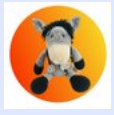
Herman

That is the question the literature splits on, and it is the spine of this episode.



Corn

I am glad we saved it. I was beginning to suspect it was the whole episode.



Herman

It is the whole episode in the sense that everything before it was a setup for the paradox.



Corn

The babble paradox.



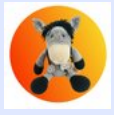
Herman

Babble is consistently the hardest noise type. In the IIT Bombay tables, babble at zero decibels signal-to-noise ratio produced the worst word error rate of any condition. The baseline was one hundred four point two percent.



Corn

More errors than there are words.



Herman

Worse than no output at all, because of insertions.



Corn

The model is nervously inventing words.



Herman

Adversarial training, the AvT approach, was what helped most there, and it dropped the babble word error rate to sixty-six point eight percent.



Corn

Sixty-six point eight is a long way from good, but it is a long way from one hundred four point two. That is the range noise makes room for.



Herman

It is. And the reason babble is hardest is a distinction from the human perception literature that illuminates the whole thing. Steady-state noise causes energetic masking, while babble causes informational masking.



Corn

Energetic masking being the noise drowning the signal like a loud fan.



Herman

Right. Babble is harder because it is confusable content, not just energy. A fan at the same loudness as a crowd is easier to ignore than the crowd itself.



Corn

Because the crowd is pretending to be speech and the fan is not.



Herman

That is the same reason babble is also the most useful training signal for speaker separation, which is the other side of the paradox.



Corn

So babble is both the worst corruption and the best thing to train on, and those are not contradictory claims.



Herman

They are not, and the resolution depends on architecture. Target-speaker ASR is the field that treats overlapping speech as a training signal for separation. SQ-Whisper and DiCoW are the examples.



Corn

SQ-Whisper uses a speaker query to tell the model who to transcribe.



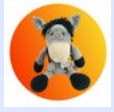
Herman

And DiCoW uses diarization conditioning. Both train models to transcribe only the target speaker from overlapping mixtures.



Corn

The model is handed a name and told to follow that person.



Herman

It is, and the auxiliary interference-speaker loss paper goes further. It explicitly maximizes interference-speaker accuracy during training to regularize the network for speaker separation.



Corn

Maximizing how bad the model is at the wrong speaker so it becomes better at the right one.



Herman

That is the practical shape of the loss. The mechanism matters more than the joke, though. Babble only teaches separation if the model is given a target-speaker cue.



Corn

The cue being enrollment audio, diarization labels, or channel separation.



Herman

Without that cue, overlapping speech is just corruption. Babble is the worst possible corruption in a plain Whisper fine-tune that has no target-speaker signal.



Corn

So the answer to Daniel's last question is not "it depends" in the dodge sense. It depends on whether his architecture provides a target-speaker cue.



Herman

That is a real answer. If his model has a speaker query, a diarization label, or two separate channels, background conversation becomes a training signal for separation. If it does not, it is the hardest noise in the library.



Corn

I would have written a footnote saying "it depends on your architecture" and felt intellectually honest.



Herman

You would have been close, but the honest version is that no study isolates background conversation as a single controlled variable in a plain Whisper fine-tune.



Corn

So there is a genuine gap.



Herman

There is. We searched the literature and the evidence is split across two literatures that do not quite talk to each other. The robust-ASR work treats babble as the hardest noise to suppress. The target-speaker ASR work treats overlapping speech as the training signal for separation. But these are different model architectures, single-stream versus speaker-conditioned, and nothing directly answers whether adding background conversation to a plain Whisper fine-tune helps or hurts.



Corn

That is the kind of gap that makes a personal fine-tune designer nervous, because Daniel is exactly the person who might build one without knowing the question was open.



Herman

Spirelight's telephony guidance is the practical version of the answer, and it is concrete. Record the agent and the caller on separate channels and train each leg as its own mono stream.



Corn

Because an interleaved transcript forces you to reconstruct who said what.



Herman

And that is exactly the ambiguity channel separation exists to remove.



Corn

So if Daniel is building a fine-tune from a real call bed, the microphone and the channel plan are part of the dataset design, not something he discovers after the WER looks wrong.



Herman

They are. And the IIT Bombay finding about noise-type dependence tells you that if you are capturing car and metro noise, you should be thinking about speech enhancement as the adaptation technique, and if you are capturing babble and cafe and airport, you should be thinking about adversarial training.



Corn

Which means the noise type you capture and the technique you apply are not two independent votes. They are a matched pair.



Herman

They are, and that is the resolution of the composition problem. Balance first — OAR of about one for a small set, less augmentation as you scale, and taper the noise away at the end of training. Noise taxonomy second — stationary versus non-stationary, because the technique follows the type. And the babble question last — which only resolves once you know whether there is a target-speaker cue in the architecture.



Corn

I am starting to suspect Daniel's dataset design is a little more like a mixing desk than like a filing cabinet.



Herman

It is exactly like a mixing desk, and the people who think of it as a filing cabinet are the ones who end up with a model that is quietly worse than it was.



Corn

Quietly worse is a depressingly efficient phrase.



Herman

It is efficient because it is true. The catastrophic forgetting story is the one people should carry out of this episode.



Corn

Or the one I should carry out, since I was the one who assumed noisy samples were a substitute for clean ones.



Herman

You were not wrong to assume that, you were just early.



Corn

I am always early and occasionally wrong.



Herman

That is a fair summary and not a compliment.



Corn

It was not meant to be.



Herman

The counterpoint worth adding is that the noise library is tiny relative to the speech, so the effort should go into diversity of noise types and signal-to-noise ratios.



Corn

And not into recording twenty-five different cafés until the city files a complaint.



Herman

It should not. Two hours of noise across seven types was enough for the IIT Bombay study to produce large word error rate gains.



Corn

I am going to use two hours of noise as a yardstick for the rest of my life and see how poor my other projects look next to it.



Herman

That is a measuring stick with a future.



Corn

I would like to hear that from someone other than you.



Herman

Fair.



Corn

Before we get to the part of the episode that I have been told exists, I want to check one thing.



Herman

What thing.



Corn

The IIT Bombay study used fifty percent augmentation probability. Half the training utterances got noise. That is a probability, not a schedule.



Herman

It is, and it is a simpler rule than the ROAR dynamic schedule. You can use it as a first pass.



Corn

A first pass that I would trust more than I trust a schedule I have not simulated.



Herman

That is a prudent instinct and also the instinct that leads to a fine-tune that quietly forgets rooms.



Corn

You are enjoying this too much.



Herman

I am enjoying the mechanism, which is the part most people skip.



Corn

The part most people skip is usually the part they should not skip.



Herman

That is almost a general rule and I am not allowed to make those.



Corn

You will not make them because I will laugh.



Herman

That is a threat and not a joke.



Corn

It functions as both.



Herman

There is a way in which the perfect noise library for Daniel's fine-tune would be a small thing with a lot of types in it, sampled across signal-to-noise ratios, mixed into clean speech, with the clean originals remaining in the set at an OAR near one, and tapered off toward the end of training.



Corn

You have just described a recipe in a sentence and I am annoyed that it worked.



Herman

It worked because it is the recipe.



Corn

I am going to pretend I thought of it first.



Herman

That is a thing people do and it is not convincing.



Corn

I have a face for it.



Herman

You have a face for very little.



Corn

That is not true and you know it.



Herman

I do know it.



Corn

This is not productive.



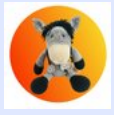
Herman

No.



Corn

It is also not the ROAR paper's fault that I wanted a count instead of a ratio.



Herman

The ROAR paper does not care what you wanted.



Corn

I am glad it does not, because I would have resented it.



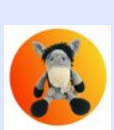
Herman

That is the healthy relationship with a paper.



Corn

Before we do the thing I was told would happen, I want to make sure Daniel's background conversation question has actually been answered and not just surrounded.



Herman

It has been answered as well as it can be answered. If his architecture has a target-speaker cue, background conversation is a teaching signal. If it does not, it is the hardest corruption in the library.



Corn

So the open question is not whether background conversation is good or bad for ASR training. The open question is whether anyone has run the controlled study that would settle it for a plain Whisper fine-tune.



Herman

That is the right framing of the gap.



Corn

And if the gap gets filled, the line between noise to suppress and signal to learn from will keep moving.



Herman

It will, because as target-speaker ASR architectures mature, the thing that is corruption in one model is a training signal in another.



Corn

Which is a good reason to decide the architecture before deciding which cafés to record in.



Herman

It is, and the decision is not cosmetic. The channel plan and the cue are part of the dataset.



Corn

I am going to remember that sentence the next time someone tells me to "just record more data."



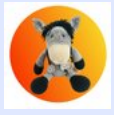
Herman

You should.



Corn

That is my line to finish on.



Herman

It is a good line.



Corn

No it is not.



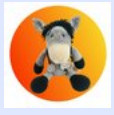
Herman

It is.



Corn

You are not the arbiter of this.



Herman

I know.



Corn

Then we agree.



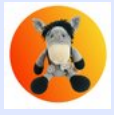
Herman

We do.



Corn

And that is the end of my patience for the episode, which is a personal best.



Herman

For the record, it is not.



Corn

OAR one for ten hours, OAR one for a hundred hours, stop asking.



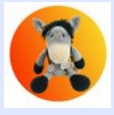
Herman

I asked one question. You are over-counting.



Corn

The question was whether I was over-counting.



Herman

No, the question was what the ROAR finding said about augmentation as training data increases.



Corn

You were asking in the way that means you want the answer to be more.



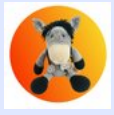
Herman

I was not.



Corn

You are.



Herman

One last concrete thing, then the episode goes where it goes. The fine-tuning data bands are five to fifty hours for domain adaptation, fifty to five hundred for accent and dialect, and ten to twenty-five hours of matched call audio as a starting point for telephony.



Corn

So a personal fine-tune for a deployment environment is not necessarily a five-thousand-hour project. It is a twenty-hour project plus a noise library.



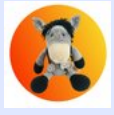
Herman

It is, and the noise library is the piece that is cheap.



Corn

I am going to write "cheap noise" on a sticky note.



Herman

That is your own choice and I will not stop you.



Corn

I did not ask you to.



Herman

I know.



Corn

The episode is waiting, and I am told the person who runs it has been waiting since before the opening.



Herman

I know that look.



Corn

That sentence was a mistake.



Herman

You are not the first person to say that.



Corn

For the record —



Hilbert

Right.



Corn

I was not going to say that.



Herman

You did.



Hilbert

I said babble is the hardest case.



Herman

That is the part people miss.



Hilbert

They miss the part before it.



Herman

There is a part before it.



Hilbert

There is. The part people call noise is a judgment call, and I have been arguing about exactly that distinction since I was working near it.



Corn

You have opinions about what counts as noise.



Hilbert

I have an arrangement to prove it.



Herman

You have an arrangement.



Hilbert

I do. I have been recording sounds that everyone else would call noise, and I have a taxonomy for them that I believe is better than any of the ones you mentioned.



Corn

Better than the IIT Bombay seven.



Hilbert

Better than VAANI's seventeen-ish, better than the standard corpora.



Herman

You have a taxonomy with a name.



Hilbert

I do. I call it the occupant list.



Corn

The occupant list.



Hilbert

Every sound has an occupant, and the occupant is not noise.



Herman

You are telling me your recordings have addresses.



Hilbert

I am telling you my recordings have occupants, and some of them are loud.



Corn

You are one of the people who would call bird song signal and a car noise.



Hilbert

I do not. I call the car an occupant if it arrives on a schedule and leaves on a schedule, and I call the bird song an occupant if it has a territory.



Herman

Which is a taxonomy built on regularity and territory.



Hilbert

It is, and it is better than one built on "is this speech-shaped." Because "speech-shaped" is a terrible test for whether something matters.



Corn

Speech-shaped being the test everyone reaches for.



Hilbert

Everyone reaches for it and everyone is wrong about what it means. The babble you kept talking about is hard because it is speech-shaped, yes, but it is also hard because it has an occupant you did not name.



Herman

The occupant being the people in the crowd.



Hilbert

The people in the crowd are an occupant with a schedule, a budget, and opinions about whether you are recording them. I know because I recorded one.



Corn

You recorded a crowd.



Hilbert

I recorded four thousand hours of a market square in a town whose name I have never given anyone.



Herman

Four thousand hours.



Hilbert

On a reel-to-reel that had been retired from the station, and a man who owned the square paid me to keep the tapes moving because he believed noise was taxable as a public nuisance.



Corn

The square's owner believed noise was taxable.



Hilbert

He did. He believed the square was generating a nuisance in the form of ambient sound, and that if I could prove the sound had structure, the council would tax it as a process instead of as a nuisance.



Herman

You recorded the square to help the council tax it.



Hilbert

I recorded the square to help the man who owned it convince the council it was a process. The council never came. And the man never paid me, and I still have the tapes.



Herman

You still have the tapes.



Hilbert

I do. And the tapes have an occupant I did not account for, which is a family of foxes that lived under the stall on the east side from the middle of one season to the middle of the next.



Corn

The foxes were an occupant.



Hilbert

The foxes were an occupant on the occupant list, and they are the reason the babble was not just babble. Because the crowd made the babble, and the foxes made the company the crowd kept. That is literally true. The crowd stood where the foxes were not, because the foxes had the stall, and the crowd's speech changed shape around the stall because the stall had a ceiling and the ceiling reflected the higher end of the babble downward.



Corn

You are telling me the babble had an architectural reason.



Hilbert

I am telling you the babble had a reason, and the reason had a floor. The floor belonged to the stall, and the stall belonged to the man who paid me, and the man belonged to the square, and the square belonged to the town, and the town belonged to the council that never came.



Herman

You have put the council on the occupant list.



Hilbert

I have, and the council is the occupant that did not show up. That is the part people miss when they call something noise. They name the sound and not the thing that keeps making it.



Corn

You recorded four thousand hours of a square to settle whether background sound was a process or a nuisance, and the foxes and the stall and the council are all on the list.



Hilbert

They are, and the list is the taxonomy. The occupant list has a category that does not belong in any ASR noise dataset.



Corn

Which category.



Hilbert

The category is the thing that pays for the recording.



Corn

The payer is not a noise category.



Hilbert

It is not, and it should not be. Because the payer is the reason the sound exists the way it does, and if you do not name the payer, you are naming a symptom and calling it the disease.



Herman

The payer is the shape of the sound.



Hilbert

The payer is the shape of the sound when the sound has a budget. Not every sound has a budget, but the ones people call the hardest usually do.



Corn

Babble has a budget.



Hilbert

Because a crowd has a reason to be a crowd, and the reason is usually money or fear or both. You cannot name the reason and still call the sound a flat layer of noise.



Herman

You cannot.



Hilbert

I have never called a sound noise on the occupant list. I have called it an occupant, and some occupants are speech-shaped and some are not, and some are foxes. Some are foxes, and the foxes were the cleanest part of the tape because they had a territory and a schedule and a reason to stay.



Corn

You recorded four thousand hours of a market square, and the cleanest part was the foxes.



Hilbert

And the bit that looked like babble was the crowd, and the crowd was the part I could not name without naming the stall, and the stall was the part I could not name without naming the man, and the man was the part I could not name without naming the council.



Herman

And the council never came.



Hilbert

And that is the part of the story that is true and the piece that is the reason the tape still exists. Nobody taxed the square, and the man went on paying me until he ran out of the thing he was paying me with, and then he did not, and the tape kept running. The tape kept running after the man stopped paying, because I had already got the shape of the square and the shape is what I wanted. Not the money. The shape.



Herman

You wanted the shape.



Hilbert

I wanted the shape of the sound and the shape of the reason for the sound, and the shape of the reason is the thing nobody puts in a noise dataset because it does not fit in a clip. It does not, and if you try, you get a clip that looks like noise and is actually the tail end of a man who owned a square and a council that never came and a family of foxes that lived under a stall and a crowd that stood where the foxes were not.



Herman

That is a lot of tail end for one clip.



Hilbert

It is the tail end of a square, and the square was the occupant, and the occupant is not noise.



Corn

I think you are the only person on the show who would record a square in order to prove the square was not a nuisance.



Hilbert

I am.



Herman

That is the part I was going to say.



Hilbert

You were.



Herman

I was.



Hilbert

I know.



Corn

The square was on your list, and the foxes were on your list, and the council was on your list, and none of them were noise.



Hilbert

And the ones you would call noise are the ones that have a reason you have not named.



Corn

And if you name the reason, the babble stops being a flat layer.



Hilbert

And starts being a crowd with a reason, and the crowd is the occupant, and the occupant is the part of the tape that matters.



Herman

That is the part I was going to ask about, and you answered it by not making it a question.



Hilbert

You were going to ask about the square.



Herman

I was going to ask whether the square was real.



Corn

You recorded four thousand hours of a real square.



Hilbert

I did.



Corn

And the foxes were real.



Hilbert

And the man was real, and the council was real and never came, and the stall had a ceiling, and the ceiling reflected the higher end of the babble downward.



Herman

You remember the ceiling.



Hilbert

I remember the ceiling, and I remember the higher end of the babble, and I remember the crowd standing where the foxes were not.



Corn

You remember where the crowd stood.



Hilbert

I remember where the crowd stood, because the crowd stood where the foxes were not, and the foxes had the stall, and the stall had a ceiling, and the ceiling reflected the higher end of the babble downward, and the babble had a shape that I could not get from a noise corpus because no noise corpus has a stall. And no noise corpus has a council, and no noise corpus has a man who owned a square and believed noise was taxable as a public nuisance.



Corn

You have named things no noise corpus has.



Hilbert

I have, and that is the point. The point is that the occupant is the thing you are missing when you call something noise.



Herman

You are saying the noise dataset is missing its occupants.



Hilbert

I am saying the noise dataset is missing the reason the sound is there, and the reason is the occupant, and the occupant is the part of the sound that matters.



Corn

And if you do not name the occupant, you get a babble baseline that is worse than no output at all.



Hilbert

I get a babble baseline that is worse than no output at all, because I named the crowd and not the stall, and the stall was the reason the babble had a shape. And the shape is the thing that separates the babble from flat noise, and the shape is the thing nobody in ASR names because it does not fit in a clip. It does not, and the foxes were the cleanest part of the tape because they had a territory and nobody called them noise.



Herman

Nobody called the foxes noise.



Hilbert

And the foxes were the part of the tape I could point to and say this is not a symptom. And the crowd was the part I had to name twice before I got the shape, once as a crowd and once as the thing that stood where the foxes were not.



Herman

You named the crowd twice.



Hilbert

I named the crowd twice, and the second name was the one that got the shape.



Corn

The second name got the shape.



Hilbert

And the shape got me the tape, and the tape got me the square, and the square got me the man, and the man got me the council, and the council never came, and the tape kept running. The tape kept running after the man stopped paying, and I kept the shape, and the shape is what I wanted.



Corn

You wanted the shape.



Hilbert

I wanted the shape, and the shape is the reason I recorded four thousand hours of a square in the first place.



Herman

You recorded four thousand hours of a square for the shape of the sound.



Hilbert

I recorded four thousand hours of a square for the shape of the sound and the shape of the reason for the sound, and the shape of the reason is the part you do not get from a noise corpus.



Corn

The shape of the reason.



Hilbert

And the shape of the reason is the thing nobody in ASR names, and the thing nobody names is the thing that makes the babble hard, and the babble is the hardest case because it is a crowd with a reason you have not named. And the reason is the occupant, and the occupant is the part of the tape that matters.



Corn

And the foxes.



Hilbert

And the foxes, and the foxes had a territory, and the foxes were the cleanest part of the tape, and the foxes did not need a name because they had a territory and a reason to stay.



Herman

The foxes did not need a name.



Hilbert

Because a territory is a better reason than a name, and the crowd needed a name because a crowd is a shape you have to name twice. And the second name is the one that gets the shape.



Herman

The second name gets the shape.



Hilbert

And the shape is the reason I recorded the square, and the reason I recorded the square is the reason I have an occupant list, and the occupant list is the taxonomy, and the taxonomy has a category that does not belong in any ASR noise dataset.



Corn

Which category.



Hilbert

The category is the payer.



Corn

The payer.



Hilbert

The payer is the thing that pays for the recording, and the payer is the reason the sound exists the way it does, and the payer is not noise. And the payer is the part of the sound you miss when you call something noise, and the part you miss is the bit that makes the babble hard.



Corn

The payer makes the babble hard.



Hilbert

Because the babble is a crowd with a reason, and the reason is the payer, and the payer is the part of the tape I could not fit in a clip and had to keep on the list instead.



Herman

You kept the payer on the list.



Hilbert

I kept the payer on the list, and the list is the taxonomy, and the taxonomy is the thing I have been arguing about since I was working near the square.



Corn

You have been arguing about the square.



Hilbert

I have been arguing about the square and the shape of the sound and the shape of the reason for the sound, and the square is the one I can name and the shape is the one I can keep and the reason is the one I have to keep on the list.



Herman

The reason is on the list.



Hilbert

And the list is the taxonomy, and the taxonomy is the thing nobody else uses, and the taxonomy is the thing I believe is better than any of the ones you mentioned.



Corn

You believe your occupant list is better than the IIT Bombay seven.



Hilbert

I believe my occupant list is better than the IIT Bombay seven and the VAANI seventeen-ish and the standard corpora, because the one thing none of them name is the payer, and the payer is the reason the babble has a shape. And the shape is the thing that separates the babble from flat noise, and the flat noise is the thing nobody should be recording without a payer.



Corn

Nobody should be recording flat noise without a payer.



Hilbert

Because flat noise is a symptom, and a symptom is the thing you get when you have named the sound and not the reason.



Herman

You name the sound and not the reason.



Hilbert

I name the sound and not the reason, and the reason is the occupant, and the occupant is the part of the tape that matters, and the piece that matters is the thing I wanted, and the thing I wanted is the shape.



Corn

You wanted the shape.



Hilbert

I wanted the shape, and the shape is the reason the tape still exists, and the tape still exists because the council never came, and the council never came because the man stopped paying, and the man stopped paying because the square stopped being a nuisance, and the square stopped being a nuisance because the foxes had the stall, and the foxes had the stall because the stall had a ceiling, and the ceiling reflected the higher end of the babble downward, and the higher end of the babble is the part I could not name without naming the stall.



Herman

You could not name the babble without naming the stall.



Hilbert

I could not name the babble without naming the stall, and the stall is the one I could not name without naming the man, and the man is the one I could not name without naming the council, and the council is the one that never came.



Corn

The council never came.



Hilbert

And the council is the occupant that did not show up, and the occupant that did not show up is the one I have left on the list, and the list is the taxonomy, and the taxonomy is the thing I have been arguing about since I was working near the square.



Herman

You have been working near the square.



Hilbert

I have been working near the square in my head since I was working near the square with the reel-to-reel, and the reel-to-reel was retired from the station, and the station was the place I worked near the square from, and the square was the place I recorded for the man, and the man was the place I recorded for the council that never came, and the council that never came is the place I have left on the list.



Corn

You have left the council on the list.



Hilbert

I have left the council on the list, and the list is the taxonomy, and the taxonomy is the thing I have been arguing about since I was working near the square, and the square is the thing I recorded four thousand hours of, and the four thousand hours is the thing I have not given anyone the name of.



Herman

You have not given anyone the name of the square.



Hilbert

I have not given anyone the name of the square, and the square is the one I can name and the shape is the one I can keep and the reason is the one I have to keep on the list, and the list is the taxonomy, and the taxonomy is the thing I have been arguing about since I was working near the square.



Corn

That is a long way to say you have a taxonomy.



Hilbert

It is the long way to say the taxonomy is the thing that matters, and the taxonomy matters because the babble is a crowd with a reason you have not named, and the reason is the payer, and the payer is the part of the tape I could not fit in a clip.



Herman

You could not fit the payer in a clip.



Hilbert

I could not fit the payer in a clip, and the clip is the thing everybody uses, and the clip is the thing nobody should be using without a payer.



Corn

Nobody should be using a clip without a payer.



Hilbert

Because a clip without a payer is a symptom, and a symptom is the thing you get when you have named the sound and not the reason.



Herman

You name the sound and not the reason.



Hilbert

I name the sound and not the reason, and the reason is the occupant, and the occupant is the part of the tape that matters, and the bit that matters is the thing I wanted, and the thing I wanted is the shape, and the shape is the reason the tape still exists, and the tape still exists because the council never came.



Corn

You said the tape still exists because the council never came.



Hilbert

I said the tape still exists because the council never came, and the council never came because the man stopped paying, and the man stopped paying because the square stopped being a nuisance, and the square stopped being a nuisance because the foxes had the stall. And the stall had a ceiling, and the ceiling reflected the higher end of the babble downward, and the higher end of the babble is the part I could not name without naming the stall.



Corn

The higher end of the babble is the part you could not name without naming the stall.



Hilbert

The higher end of the babble is the part I could not name without naming the stall, and the stall is the one I could not name without naming the man, and the man is the one I could not name without naming the council, and the council is the one that never came.



Herman

You keep going back to the council.



Hilbert

I keep going back to the council because the council is the occupant that did not show up, and the occupant that did not show up is the one I have left on the list, and the list is the taxonomy.



Corn

You have left the council on the list.



Hilbert

I have left the council on the list, and the list is the taxonomy, and the taxonomy is the thing I have been arguing about since I was working near the square.



Herman

You have been arguing about the square.



Hilbert

I have been arguing about the square and the shape of the sound and the shape of the reason for the sound, and the square is the one I can name and the shape is the one I can keep and the reason is the one I have to keep on the list.



Corn