Episode Book

EPISODE #6299

Finding Recurring Urban Sounds With Four Examples

PUBLISHED	27 October 2026 at 12:34 UTC
DURATION	24:35
LISTEN	myweirdprompts.com/episode/few-shot-sound-event-detection-73948/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:52 UTC

SYNOPSIS

Training an audio detector does not always require hundreds of manual annotations. Borrowing methods from bioacoustic few-shot sound event detection, a system can use pretrained audio neural networks to locate recurring noises like train horns from just four or five distinct examples. The discussion covers model-assisted labeling workflows, using prototypical networks on top of PANNs embeddings, avoiding background acoustic leakage across recording sessions, and balancing precision against recall when reviewing candidate audio segments.

TRANSCRIPT



Corn

Recording the noise is easy. Remembering which noise happened where is the unpaid second shift. Daniel was collecting voice notes for his ASR fine-tune and noticed the recurring soundtrack around him, lottery stands playing music, cars honking, the train passing. He wants recording and annotation to remain separate jobs. Could marking four train horns give him a detector that finds the rest, with its guesses going through his review before joining the training set?



Herman

Four is close to the operating point of an actual benchmark. Few-shot sound event detection. DCASE has a task where you give a system five annotated animal calls and ask it to locate more of that sound in long recordings.



Corn

Birds have already negotiated the annotation budget.



Herman

Expert time drove that design. Hours of recordings, very few people qualified to mark them, and an awful lot of empty listening. Daniel has a similar bottleneck. He knows what he wants marked, but repeatedly scrubbing through his own recordings is expensive.



Corn

Especially when he already knows what the train sounds like. The computer is the new employee here.



Herman

Its induction can be surprisingly short. Its probation should be longer.



Corn

I can hear you wanting to abolish probation.



Herman

I want a first pass worth reviewing. Different ambition. The benchmark is called Few-shot Bioacoustic Event Detection, and the important part is that the supplied examples have start and end times. The model has to find occurrences, not merely announce that an entire recording contains a bird.



Corn

That distinction matters for voice notes. A tag saying train somewhere in these twelve minutes leaves almost all the work intact.



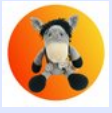
Herman

Detection gives you regions to inspect. In the DCASE twenty twenty-four evaluation, the prototypical-network baseline got an event-based F-score of forty-one point six. The strongest team system reached sixty-five point two. Those combine precision and recall, so neither means that percentage of its suggestions will be correct.



Corn

Nor that it has found that percentage of the horns.



Herman

Yes. Well... yes to your objection, not to either interpretation. Precision concerns how many proposed events are real. Recall concerns how many real events it finds. Event scoring also cares about locating them acceptably. You can improve one while harming another by changing the detection threshold.



Corn

If I make it nervous enough, everything becomes a horn.



Herman

Lowering the threshold can increase recall and flood the queue. Raising it can create a lovely clean queue while silently missing most events. Corn, which would you tolerate for your own recordings?



Corn

Extra suggestions, within reason. I can dismiss a passing car quickly. I cannot correct a missing horn if the interface never takes me there. However, if it marks every passing car, I have purchased a traffic diary.



Herman

That sets the objective more usefully than maximizing a benchmark score. Time saved, plus enough coverage that the marks remain useful. Five examples can produce that sort of noisy first pass. The benchmark supports trying it, not transferring its score straight onto Daniel's street.



Corn

The lottery stand is probably worse than the horn. Music changes. A car parks in front of it. Somebody talks over it. Does he mean that particular stand, or any music coming from a stand?



Herman

Those are different target definitions. A loud tonal horn feels like a promising first class. Music under speech is a harder boundary. I would begin with the horn, without promising that every urban event will behave similarly.



Corn

Daniel also wants a name for the journey from doing everything himself to supervising suggestions and eventually checking occasionally. I would call the middle bit employment.



Herman

Model-assisted labelling is the useful search term. Model-in-the-loop labelling and pre-labelling also describe it. For the whole evolving arrangement, people say data flywheel, data-centric learning loop, or continuous learning loop. No single canonical academic name captures those phases.



Corn

Continuous learning sounds like it updates itself whenever I leave the room. I would rather the dataset change first, then a deliberate training run.



Herman

Sensible distinction. The name doesn't specify the deployment policy. You can collect corrections all week, train a candidate model, compare it with the existing one, then decide whether to replace it. Nothing requires changing the detector during a recording session.



Corn

Separate recording, separate annotation, separate decision to trust the new model. That suits the actual work better than one endlessly moving system.



Herman

Underneath, two terms matter. Active learning selects which examples would be valuable for a human to annotate. Self-training takes the model's predictions as pseudo-labels for further training. You can combine them, but they do different jobs.



Corn

Accepting suggestions by hand sounds less like pure self-training and more like ordinary supervised training with cheaper label entry.



Herman

It is. Good correction. Human-reviewed labels support supervised retraining. Calling every assisted workflow self-training blurs that distinction. Self-training applies when predictions themselves become training targets without equivalent human verification.



Corn

I would like my labels to retain that distinction after export.



Herman

Keep their provenance. Human-marked, human-corrected, model-only. Confidence and model version are useful alongside them. Otherwise next month's experiment may treat a detector's guesses as independent evidence that the same detector works.



Corn

A reference written by the applicant.



Herman

Silver is familiar vocabulary in weak supervision and programmatic labelling. Snorkel is a prominent example, combining imperfect labelling functions and accounting for their errors and correlations. Those functions can be heuristics rather than trained classifiers. It doesn't make every silver-label workflow a Snorkel workflow.



Corn

A rule saying sustained tonal energy might flag horns, then. Useful, but also capable of discovering an extremely confident refrigerator.



Herman

Background negatives help distinguish those. Four positives alone don't define everything the target isn't. A practical detector also needs ordinary street sound, speech, car honks, and whatever repeatedly fools it as comparison material.



Corn

You are about to tell me four clips become four vectors and everything gets elegantly small.



Herman

You know my tells.



Corn

You lean forward when the classifier fits on a napkin.



Herman

The pretrained encoder doesn't fit on the napkin. That is where much of the learned acoustic knowledge lives. It turns each example into an embedding, a vector representing acoustic characteristics. A prototype classifier averages the target examples' vectors. New audio windows are compared with that average, usually alongside some representation of background.



Corn

Four examples customize an existing listener. They don't teach a newborn model what sound is.



Herman

Precisely the distinction... sorry, that sounded like a certificate. Most of the expensive learning happened before Daniel supplied his clips. His examples define the local target. They may not require updating the encoder's weights at all.



Corn

Does averaging throw away the useful difference between a distant horn and a close one?



Herman

It can. The hope is that the embedding preserves their shared identity while tolerating irrelevant differences. With very few examples, the mean may land somewhere unrepresentative. Multiple prototypes or a different lightweight classifier can help later, but first you need to see where this simple approach breaks.



Corn

One example gives you no way to distinguish the horn from its accidental companions. Four from separate encounters at least begin to expose those companions.



Herman

That is why one-shot performance can deteriorate so badly. One support point gives no within-class variation. Four helps, provided they really are separate examples. Four slices of the same blast are not four independent acoustic situations.



Corn

The train has submitted the same reference four times.



Herman

Nolasco and colleagues titled their paper Learning to detect an animal sound from five examples. They demonstrated detection even for sound categories unknown during the algorithm's training. Five is an established experimental setting, rather than a number someone picked for a demo.



Corn

Still not a minimum requirement stamped on the box.



Herman

No. Four is enough to attempt a useful prototype. Five varied examples is a well-supported starting point. Neither establishes readiness for autonomous operation. I can't give Daniel a published urban-horn sample count that guarantees a reliable recognizer.



Corn

I would prefer five encounters over fifty fragments from one crossing. Though I might deliberately keep one difficult, distant example out of training. Otherwise we congratulate the system for recognizing the recording it has effectively already met.



Herman

Would you hold out a whole recording session, rather than scattered clips?



Corn

Yes. The background can identify a session almost as well as the target. If everything from one walk leaks into both sets, we may be testing whether it recognizes that walk.



Herman

That changes the first experiment. Use some manually checked audio from another session as an evaluation set. Daniel called the seed examples held-out gold samples, but once they establish the prototype, they are support or training examples. Truly held-out examples stay outside that process.



Corn

Four seeds, separate test audio, and permission to discover it's only good at one horn. I can work with that.



Herman

One recurring horn, heard at different distances, would be my simplest initial case. Then inspect the confusions before widening the class. The DCASE results vary enormously across sound groups. A respectable average can conceal a category that barely works.



Corn

What I do not want is an afternoon of choosing between six audio backbones before I have one usable suggestion.



Herman

Start with PANNs. Pretrained Audio Neural Networks, trained on AudioSet, covering hundreds of sound classes. The Python package is called panns-inference. Use its embeddings as the representation, then add the small classifier.



Corn

Small classifier. Large assembly instructions.



Herman

There isn't a single established package I can confidently recommend that takes four arbitrary examples and delivers this complete reviewable detector. The realistic stack has three pieces. An embedding extractor, a detection or prototype component, and an annotation interface.



Corn

You install all three before breakfast and ask me to use them after lunch.



Herman

Only once they run.



Corn

That qualification has consumed several lunches.



Herman

You would find the broken region selector before I did. Which is why the interface belongs in the first prototype, rather than after the exciting model work.



Herman

BEATs and AST are alternative pretrained backbones. A project called voice classification bundles those alongside PANNs and other encoders. That can reduce comparison plumbing. OpenL3 also produces audio embeddings, but its main repository hasn't been updated since twenty twenty-three, so I wouldn't make it the default for a fresh setup.



Corn

PANNs first, keep the alternatives available. The fastest comparison is one I don't run yet.



Herman

For the detector, the DCASE twenty twenty-four Task Five baseline provides a working prototypical-network implementation. It is bioacoustic code, so expect adaptation around audio segmentation, durations, and input formats. It is a reference implementation, not a finished street-audio application.



Corn

The horn has declined to vocalize like a mammal.



Herman

Duration really does matter. A short honk and a long blast interact differently with window size. Windows that are too long mix the event with lots of unrelated audio. Windows that are too short may fragment one event into repeated detections. You need a step that joins adjacent positive windows into reviewable regions.



Corn

Otherwise one train becomes twelve tasks.



Herman

Label Studio can handle the gold annotation and the review interface. Import the detector's proposals as pre-filled time regions. Daniel listens, accepts a region, adjusts its boundaries, changes its class, or rejects it.



Corn

The model's detections would be silver labels that I promote through review.
Ivan: Silver Kestrel. Forty one, seven three. Two two, nine zero. The train horn is not the train. The vinegar jar is sealed twice. Forty one, seven three. The onions are counted under the lid. Nightjar closes.



Herman

Copper Heron. Receipt confirmed.



Corn

Herman, what on earth was that?



Herman

Reviewed regions can enter the gold training set. Rejections matter too, as hard negatives, provided the surrounding audio has actually been checked. An unmarked section of a partly reviewed file cannot automatically mean no horn.



Corn

Hm.



Corn

An accept button also needs to mean something definite. Did I hear a horn, or did I check both boundaries and every event nearby? Those are different amounts of work.



Herman

Define that in the annotation policy. A confirmed positive region doesn't certify the entire file. This becomes especially important when retraining on long recordings, where missing annotations can accidentally teach the model to suppress real events.



Corn

I would keep the clip playing slightly before and after the proposal. Tight cropping can hide the very clue I need to reject it.



Herman

Context without forcing full playback. Good. Corn, would you spend your limited review time on the uncertain proposals or the cleanest ones?



Corn

Neither exclusively. Clean ones get me an initial useful dataset. Uncertain ones help repair the boundary. I also want some stretches where it proposes nothing, because the queue cannot tell me what it never noticed.



Herman

That last category prevents a particularly nasty feedback loop. Suppose all the initial positives came from one crossing. The detector finds more recordings of that crossing's horn. You approve them. Its dataset grows, but its definition barely broadens.



Corn

More confident. No better travelled.



Herman

Confidence alone won't expose confidently missed events. Active learning needs some coverage of the wider collection, not just uncertainty scores. Disagreement between models, acoustic diversity, and random checks can contribute different information.



Corn

How much of the promised time saving survives that wider search? I can imagine a detector making annotation faster while quietly giving me a second job auditing the detector.



Herman

The closest urban analogue is SONYC, the New York urban-noise dataset. Online active learning there reduced annotation effort by a factor of five. That's a meaningful reason to try the loop, though it doesn't promise Daniel the same saving.



Corn

Five times less listening would justify some plumbing.



Herman

On rare-event audio, another active-learning experiment achieved performance similar to fully labelled training while annotating only two percent of the data. The important condition is rare events. Most randomly selected listening contains nothing useful, so selecting informative material buys a great deal.



Corn

That is not permission to label two percent of any folder and retire.



Herman

Correct. Different tasks, different budgets. A later bioacoustic comparison reached mean average precision in the high sixties or low seventies against seventy-five for full supervision, using a small fraction of the annotations. Close, not identical, and a different



Herman (continued)

metric from event F-score.



Corn

Those leftover errors might be exactly the distant horns Daniel cares about. Average performance doesn't tell him which recordings lose.



Herman

Nor has selection strategy been settled. A recent challenge to mismatch-first farthest-traversal argues that it can choose acoustically redundant segments under tight budgets. The proposed alternative emphasizes covering the collection while using disagreement more gently.



Corn

A model can find twenty puzzling clips of the same delivery van and call that curiosity.



Herman

Yes. No... let me narrow that. A selection method can overvalue repeated versions of one puzzle. The practical question is whether another reviewed example contributes something new. Pure uncertainty ranking doesn't guarantee that.



Corn

Daniel's separation of recording and annotation helps here. He can finish collecting normally, then look across the batch instead of stopping every time the detector becomes anxious.



Herman

Batch selection gives you that opportunity. Keep recording timestamps and session identities. Inspect coverage across conditions, not merely the highest scores. Different distances, microphone orientations, traffic backgrounds, and event lengths can all matter.



Corn

You sound ready to prescribe six more walks.



Herman

I'd inspect the existing collection first. You nearly accused me of exercise.



Corn

I withdraw the serious charge.



Herman

Domain shift is the technical concern. A horn recorded nearby on a quiet street can move to a different part of the embedding space when heard through traffic from another platform. The sound source may be identical while its recorded appearance changes substantially.



Corn

That makes the question of how many examples partly a question of which examples. Four useful encounters can beat a much larger set that never leaves one acoustic situation.



Herman

Provided the target itself remains coherent. If there are different horn types, they may need broader support or separate prototypes. We shouldn't assume duration differences are merely distance effects.



Hilbert

...and the committee still charged me seventeen pounds for submitting the long blast in the short-blast column. That was the Railway Listening Society. Late eighties. I was doing tape duplication then. Their minutes had disputed those columns for eleven years. Nobody disputed the fee.



Corn

Administrative certainty travelled further than the horn.



Hilbert

Two horns at that crossing, Herman. Short platform blast, longer crossing blast. In my log the longer one reached the back room on damp evenings. The short one didn't. That was my window and my notebook, not a rule about railways. Your single-horn starting case needs checking. Four examples could help, but four short blasts would have left half my entries out.



Herman

I treated the recurring source as more uniform than we know it is. Different blast types are a sensible distinction to check.



Hilbert

I marked every horn from the flat for nine months. Handwritten times. Blue meant short except on committee days, when blue meant awaiting a ruling. Don't ask for the colour key. There were forty-three colours and the secretary required a separate pencil for each minute of the hour. Cut my cough before colours. Anyway, I cross-checked every entry against a timetable I also wrote by hand. Never matched once. The timetable was for the other side of the station. They wouldn't let me amend it without another seventeen pounds.



Corn

You could pay again or continue documenting the wrong platform.



Hilbert

I continued. Corn, you were careful earlier about what an accept button certifies. I wanted the same care from the secretary. I offered to send the log with the two blast types marked in words. No colours. He wanted it recopied onto the society form first. That took three evenings. He was decent about reading it once it arrived.



Herman

...The event definitions need to survive transfer between tools. Short and long blasts could remain subtypes even if Daniel ultimately wants one train-horn category. That lets him check whether improvement on one conceals persistent misses on the other.



Corn

It also lets him change his mind later without replaying the entire collection. Keep the useful detail when reviewing, then combine categories for the detector if that helps.



Herman

The common misconception is that accumulating approved detections will eventually cover the whole class by itself.



Corn

It covers what the detector already finds. Broader coverage needs checks outside its suggestions, including other sessions and other horn types. Silver labels save effort only if reviewing them doesn't become a ceremony of agreement.



Herman

Whether the loop earns less supervision should depend on performance on separate, manually checked recordings. Four or five examples can start the experiment. They don't decide its final level of autonomy.



Corn

I would watch whether each new review batch saves more time than it costs, including the evenings spent correcting formats. Better than scrolling is a useful first ambition.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode seven, Building Custom ASR Tools; and episode eight, Building Your Own Whisper. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave us a review wherever you listen.



Corn

See you tomorrow.

SOURCES

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

1. **DCASE 2024 Task 5 task description (5-shot setup) [primary]**
   https://dcase.community/challenge2024/task-few-shot-bioacoustic-event-detection

2. **DCASE 2024 Task 5 full results tables [primary]**
   https://dcase.community/challenge2024/task-few-shot-bioacoustic-event-detection-results

3. **PANNs repo [primary]**
   https://github.com/qiuqiangkong/audioset_tagging_cnn

4. **panns-inference Python package [primary]**
   https://github.com/qiuqiangkong/panns_inference

5. **DCASE 2024 Task 5 baseline code [primary]**
   https://github.com/c4dm/dcase-few-shot-bioacoustic/tree/main/baselines/dcase2024_task5

6. **prototypical network baseline explanation [primary]**
   https://github.com/c4dm/dcase-few-shot-bioacoustic/tree/main/baselines/deep_learning

7. **Nolasco et al., Learning to detect an animal sound from five examples, 2023-05-22**
   https://arxiv.org/abs/2305.13210v1

8. **Nolasco et al., Few-shot bioacoustic event detection at the DCASE 2022 challenge, 2022-07-14**
   https://arxiv.org/abs/2207.07911v1

9. **Yang et al., A Mutual learning framework for Few-shot Sound Event Detection, ICASSP 2022**
   https://arxiv.org/abs/2110.04474v2

10. **Bhosale et al., Semi Supervised Learning For Few-shot Audio Classification By Episodic Triplet Mining, 2021-02-16**
https://arxiv.org/abs/2102.08074v1

11. **Zhao, Heittola & Virtanen, Active Learning for Sound Event Detection, rev. 2020-09-09**
https://arxiv.org/abs/2002.05033v2

12. **Zhang & Virtanen, Hybrid Disagreement-Diversity Active Learning for Bioacoustic SED, EUSIPCO 2025**
https://arxiv.org/abs/2505.20956v2

13. **Lindholm et al., Aggregation Strategies for Efficient Annotation of Bioacoustic Sound Events Using Active Learning, 2025-03-04**
https://arxiv.org/abs/2503.02422v1

14. **Lindsey et al., Online Active Learning For Sound Event Detection, ICASSP 2024**
https://arxiv.org/abs/2309.14460v1

15. **Zhang & Virtanen, Cover First, Disagree Softly, rev. 2026-09-18, DCASE Workshop 2026**
https://arxiv.org/abs/2607.13571v2

16. **Liu, Chen & Zhang, Power Pooling Operators and Confidence Learning for Semi-Supervised SED, 2020-05-23**
https://arxiv.org/abs/2005.11459v1

17. **Liang et al., Mind the Domain Gap: a Systematic Analysis on Bioacoustic SED, 2024-03-27**
https://arxiv.org/abs/2403.18638v1

18. **A Survey on Programmatic Weak Supervision**
https://arxiv.org/abs/2202.05433

19. **Ratner et al., Snorkel: rapid training data creation with weak supervision, VLDB Journal 2019**
https://pmc.ncbi.nlm.nih.gov/articles/PMC7075849/

20. **Snorkel AI, Making automated data labeling a reality in modern AI**
https://snorkel.ai/blog/automated-data-labeling/

21. **OpenL3 (606, MIT, last push 2023-06-17)**
https://github.com/marl/openl3

22. **PyTorch OpenL3 port**
https://github.com/hugofloresgarcia/torchopenl3

23. **pretrained audio backbones incl. PANN, AST, BEATs (Apache-2.0, last push 2026-08-09)**
https://github.com/lkleonk/voice_classification

24. **Lightweight, NumPy-first toolkit for audio embeddings and sound-event detection (MIT, last push 2026-07-26, 1)**
https://github.com/hazelpe/resona

25. **browser-based PANNs SED via ONNX**
https://github.com/TUT-ARG/WebSED