Episode Book

EPISODE #6726

Why Audio Augmentation Cuts Out What Matters

PUBLISHED	21 November 2026 at 16:27 UTC
DURATION	30:30
LISTEN	myweirdprompts.com/episode/audio-data-augmentation-techniques-15255/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.36
SOURCES	13 cited by the research (listed at the end)
BOOK GENERATED	21 November 2026 at 19:31 UTC

SYNOPSIS

Training speech models on clean studio recordings leaves them fragile when real life gets loud. This breakdown explores the four core audio augmentation interventions: speed perturbation, SpecAugment, additive noise, and room impulse response convolution. It examines why masking features forces models to generalize while time warping barely changes error rates, how augmentation turns overfitting into underfitting, and why standard datasets like MUSAN contain surprisingly little genuine noise.

TRANSCRIPT



Corn

The gap we keep circling back to on this show is the same one every time: you train on clean audio, and the world comes to you loud. Daniel's prompt today sits right in that gap — he wants the standard augmentation toolbox explained, but that's only the warm-up. What he actually wants is the freehand version. Instead of downloading a generic noise corpus, you record the noise you personally face — traffic, a coffee shop, his son Ezra screaming and throwing things around the apartment — and mix those sounds into clean training clips the way film people do Foley. He's tried capturing these moments and they don't cooperate, because babies don't scream on cue. His questions are practical: can you build a personal sound library, how do you mix it well, and where does it stop helping.



Herman

And it's a good week for it, because the literature on this is lopsided in a really interesting way. The standard techniques are documented to death. The personal version — nothing. Nobody has published on recording your own household and using it for augmentation.



Corn

Which means the second half of this episode is us reasoning in public.



Herman

My favourite genre.



Corn

Before we get to the baby, though, the four standard techniques — because Daniel's right that it sounds like one task, and it isn't. It's four tasks happening at four different places in the pipeline.



Herman

That's the frame I'd want everyone to hold onto. Speed perturbation touches the waveform. SpecAugment touches the features — the log mel spectrogram, after the audio has already been turned into an image. Additive noise is a mixing operation, two signals summed. Reverb is a convolution. Four different operations, four different points of intervention, and each one can go wrong in its own particular way.



Corn

And that's not just taxonomy for its own sake, right? The location matters practically.



Herman

It matters enormously. If you apply speed perturbation, you have to do it before you compute anything else — you're changing the raw samples. If you apply SpecAugment, you're doing it after feature extraction, which means it happens inside the training loop, on tensors, not on audio files. And if you get the order confused — say you mask a spectrogram and then try to treat it like a waveform problem — nothing explodes, it just quietly does nothing sensible. The failure mode of doing augmentation in the wrong place is silence, not error messages.



Corn

Walk me through them. Start with speed, because I think people conflate two things there.



Herman

They do. Speed perturbation resamples the waveform, and that changes tempo and pitch together — a zero point nine speed factor gives you a slower talker who is also a lower-pitched talker, formants shifted down and everything. Tempo perturbation is the fancier version that changes the rate while preserving the pitch. In practice, the standard recipe is speed perturbation at around zero point nine and one point one, and the crude version survives because it works — you've just doubled or tripled your effective corpus with speakers who sound slightly different.



Corn

A poor man's voice cast.



Herman

And there's a reason the crude version survives, beyond just tradition. The pitch shift that comes along for the ride is itself a kind of regularization — the model can't lock onto absolute pitch as a speaker cue, because its "same speaker" now exists at multiple pitches. The impurity is the point.



Corn

Then SpecAugment, which is the famous one, from Park and colleagues at Google in twenty nineteen. It looks at the log mel spectrogram and deforms it like an image. Three moves: time warping, which bends the time axis; frequency masking, which blanks out a band of mel channels; time masking, which blanks out a stretch of time steps. The masking just sets those values to zero — and because the spectrograms are normalized to zero mean, zeroing a region is the same as setting it to the average. It's hiding information from the model so it can't memorize one exact spectral fingerprint per word.



Corn

And I want to make sure I actually understand the mechanism there, because "it hides information" can sound like it just degrades the input. How is that different from just adding static to the audio?



Herman

Static is additive — the original information is still in there, just buried. Masking is destructive — the information in that band is gone, unrecoverable, and the model has to reconstruct the word from what remains. It's the difference between reading a page through fog and reading a page with a strip torn off. In the fog you can squint. With the strip gone, you have to know the language — which words can lose their top half and still be identifiable. That's the knowledge you're forcing the model to build.



Corn

And the noise and reverb?



Herman

Additive noise is the obvious one — take your clean clip, take a noise clip, sum them with a weighting that hits a target signal-to-noise ratio. The standard corpus is MUSAN, which is about a hundred and nine hours total — but here's the catch people miss: only about six hours of that is actually noise. Nine hundred twenty-nine files. The rest is speech and music, because MUSAN was built for music-speech discrimination, and noise augmentation was an afterthought. And reverb is convolution with a room impulse response — you take anechoic speech and mathematically push it through a simulated or measured room.



Corn

So when someone says "I'm doing data augmentation," the honest answer is a question — which of four things are you doing.



Herman

And the deeper claim, the one that anchors everything today, is from the SpecAugment paper itself. Augmentation converts an overfitting problem into an underfitting problem. Your training loss on the training set goes up. On the data the model has literally memorized before, it now performs worse — and that's the intended effect, not a bug.



Corn

That's counterintuitive enough that I want to sit on it for a second.



Herman

Please do, because it has a corollary that ruins a lot of weekends. The published gains came from applying a harsh policy and then paying for it — wider networks, deeper networks, longer training schedules — to claw back the underfitting. If you turn up the augmentation and don't also turn up capacity and patience, you just get a worse model. Slowly.



Corn

A sloth could have told them that. Everything worth having costs a bigger model and a longer schedule.



Herman

Some of us train fast and regret it later.



Corn

Some of us nap through the regret. Okay — the ablation. You said time warping is one of the three SpecAugment moves, and I remember the numbers being unkind to it.



Herman

Unkind is generous. Their own Table 6 — full policy gets 10.0 on test-other. Turn off time warping entirely: 10.1. A tenth of a point. Turn off frequency masking: 11.0. Turn off time masking: 10.9. So the two masking operations are carrying the whole thing, and time warping — the deformation everyone mentions first, the one that sounds the most technically sophisticated — is worth about as much as rounding error. The authors say it plainly: it's the most expensive and least influential of the three, and it should be the first thing dropped under any budget constraint.



Corn

The showiest move in the routine is the one the choreographer would cut first.



Herman

And I think there's a lesson in why it fails, not just that it fails. Time warping bends the time axis — but speech models with the right architecture are already somewhat tolerant of time distortion, and the other augmentations in the stack, speed perturbation especially, were already stretching and compressing time anyway. The masking moves do something nothing else in the pipeline does: they delete. Subtraction was the novel operation. Stretching was a rerun.



Corn

Which raises the headline numbers — worth having in your head for scale.



Herman

On LibriSpeech 960 hours, SpecAugment got 2.8 percent word error rate on test-clean, 6.8 on test-other, without a language model — a 22 percent relative improvement on test-other over the previous state of the art. That's the bar that made masking the default.



Corn

Which raises Daniel's framework question. He assumed you'd need one. Do you?



Herman

No, and I'll go further — torchaudio's own documentation demonstrates that you don't. Their SpecAugment example is a sequential block of three transforms — time stretch, frequency masking, time masking — wrapped in a module that chains a resample, a spectrogram, the augmentation block, and a mel scale. The whole pipeline is a handful of lines. Frequency masking under the hood is slicing a band of the spectrogram and setting it to zero. Speed perturbation is a resample. Additive noise is clean plus alpha times noise. Reverb is one convolution call.



Corn

So the mystery ingredients are a slice assignment and a multiply.



Herman

That's it. Now — what a framework does buy you is three things. Tested resamplers, so you're not introducing artifacts with a badly implemented interpolation. On-the-fly augmentation in the data loader, so you're not writing terabytes of augmented audio to disk. And reproducibility, same random draws, same result. Those are real. But the operations themselves are not deep magic, and I'd rather Daniel understood them than imported them.



Corn

He's an open source developer. He'd rather understand them than import them.



Herman

Then he's the target audience for the honest version of this: the third-party library landscape, the ready-made augmentation packages — I can't vouch for any of them. Adoption numbers, maintenance status, whether the thing people recommend this year is abandoned — I haven't seen anything I'd repeat. The claim I'll stand behind is the boring one: the primitives are in torchaudio and scipy, and they're documented.



Corn

One more piece of standard-toolbox evidence before we go freehand — the room impulse response work.



Herman

Yes, because this is the strongest case for the whole environment-matching argument, and it comes from voice activity detection. Ivry and colleagues trained detectors on nearly five million augmented utterances — anechoic originals plus versions convolved with a variety of simulated room responses — and then tested in real reverberant rooms. Twenty percent average improvement in accuracy, precision and recall across every detector and every room model, compared with training on clean audio only. And their diagnosis is the sentence I'd tattoo on this episode: the real environment is reverberant, the training data isn't, and performance significantly deteriorates as a result.



Corn

Clean studio training, reverberant testing.



Herman

The mismatch is the disease. The question is which medicine.



Herman

Now it's time for another edition of Dorothy's History of Toast.

D

Dorothy

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

D

Dorothy

In January nineteen forty-three the United States government banned sliced bread, as a wartime economy measure. People were so upset the ban was lifted within two months.

D

Dorothy

You try to make one small saving and everybody just complains until you give in, that's the way of it, isn't it.

D

Dorothy

Anyway, I've got a loaf going stale on the counter, so I'd best get to it. Bye bye, sweetheart.



Herman

Mum, that one's practically an episode prompt of its own.



Corn

Twenty-two million households, one banned convenience, total collapse of public order in eight weeks. I know whose side I'm on.



Herman

Where were we — yes, the freehand version. So the literature goes quiet exactly where Daniel wants to go. Nobody has published on a personal, self-recorded Foley library for ASR augmentation. Not a paper, not a workshop note. Every noise source in the literature is a corpus or a simulation.



Corn

Which cuts two ways. Either it's unexplored because nobody thought of it, or it's unexplored because everyone who thought of it found something better.



Herman

A fair worry. But there's a strand of evidence that sits one level down from Daniel's idea and points in his favour. Blind room impulse response estimation. There's work from twenty twenty-one — the FiNS model — and a twenty twenty-four follow-up with smart glasses, where they estimate the room response of your actual room from a few seconds of your own speech. No special measurement signal. And in the listening tests, the responses estimated from your own environment reproduced the room acoustics more convincingly than measured responses taken from other rooms of similar size.



Corn

Hold on. Your own room, estimated sloppily from a few seconds of talking, beats a carefully measured response from a similar room?



Herman

That's what they found. Your room beats a stock room. And if that holds for reverberation — the ambient character of the space — it's a decent bet it holds for discrete sound events too. Your baby beats a stock baby.



Corn

Ezra certainly would take that as a compliment. He works hard at that job.



Herman

He's sixteen months old and reportedly very good at throwing things. So — how would you actually build this? Let me lay out the recipe, and I want you to poke at it, because none of this is scripture, it's assembled from what the adjacent evidence implies plus what practitioners do.



Corn

Go.



Herman

Capture the noise as isolated events, not as ambiance — you want the clatter, not ten minutes of room tone with one clatter somewhere in it. Record at the same sample rate and with the same microphone as your clean data, because the mic's frequency response is part of the noise's identity.



Corn

That's a detail I'd have gotten wrong. I'd have recorded the Foley on my phone at forty-four point one and mixed it into sixteen kilohertz training clips.



Herman

Resampling artifacts — you'd be training the model on a version of the noise it will never actually meet. And the microphone point deserves one more beat, because it's not just sample rate. Every microphone has a frequency response — some boost the highs, some roll off the lows, some have a presence peak around the speech band. If your clean dictation is recorded on the headset mic and your Foley is recorded on the phone, the noise arrives in the mix wearing a different microphone's fingerprints than it will wear at inference. The model learns to ignore phone-flavored screams. Then it meets headset-flavored screams.



Corn

Which is the same mismatch as clean-versus-reverberant, one level down. The disease again.



Herman

Then: measure each clip's level. When you mix, draw the target signal-to-noise ratio from a range, not a fixed value — a single fixed SNR is a single point, one difficulty setting, and the model learns exactly that one.



Corn

Like training a boxer against exactly one sparring partner.



Herman

And vary the overlap. Some mixes put the noise entirely inside the speech, some entirely inside the pauses, some straddling the boundary — because the model needs to learn two different lessons. "Ignore this" when the scream rides over your words, and "this is not a word" when the clatter lands in a silence and could plausibly be transcribed.



Corn

That distinction is the sharpest thing in the recipe. A thrown block in a pause is a decoy word.



Herman

Right. And it's worth saying why the decoy case is the dangerous one. When noise overlaps speech, the model's job is extraction — pull the signal out of the mess. When noise sits in a pause, the model's job is classification — is this acoustic event a word or not? Those are different failures. The first failure sounds like a wrong transcription. The second sounds like the model confidently inventing a word that was never spoken, which is a much more alarming thing to find in your notes app at midnight.



Corn

Noted. Keep a substantial clean fraction so the model doesn't forget what unspoiled speech sounds like — and then the honest caveat.



Herman

No source establishes a recommended SNR sweep or a clean-to-augmented ratio. Not the papers, not anyone. Any specific numbers you hear from practitioners — mix at such-and-such decibels, keep thirty percent clean — that's convention and folklore, not findings.



Corn

Daniel will like that, actually. He builds things where the folklore is all there is, and the fun is testing it. He found the same thing with voice cloning — an hour of fine-tuning data lost to ten seconds of reference audio. The folk numbers were all wrong.



Herman

His whole career is folklore falsification. Okay — the limits. Where does this stop helping? Four places, and the first is arithmetic. The frame-level SpecAugment work measured augmentation's benefit across data sizes up to twenty-five thousand hours, and the effect of adding it was approximately equivalent to doubling your training data. Which sounds great until you read the next line — the benefits diminish as the data grows.



Corn

Because it's a substitute for data, and substitutes saturate.



Herman

Personal Foley manufactures a second copy of your corpus. A second copy is worth less than a new first copy. If Ezra-noise doubles your data, you've gained one Ezra-equivalent, not one world.



Corn

And Daniel's corpus is small by definition — his own dictation. So he's actually in the regime where the multiplier still has value.



Herman

He is. That's the one structural advantage of his situation. But here's the second limit, and it's the one that worries me. There's a paper from October — Ibaraki and Chiang, low-resource languages — that looked past error rates to error types. SpecAugment has the most reliable aggregate improvements, but it increases tone and palatalization errors. Do you see why?



Corn

Give me a second. Tone... rests on a single acoustic cue, in the spectrum, and masking blanks bands of the spectrum.



Herman

That's it. Tone and palatalization are each signalled by one narrow acoustic property, and they occur least frequently in the data. Mask the spectrum, and you destroy exactly the cue those phonemes depend on, disproportionately. Now translate that to Daniel's case. If Ezra's scream has energy concentrated in the frequency band that carries a distinction his recognizer cares about —



Corn

Then targeted noise is worse than generic noise. The MUSAN noise at least spreads its damage around; your Foley aims it.



Herman

The noise you chose for being realistic is realistic in the worst possible place. And English doesn't have lexical tone, so he's partially shielded — but vowel contrasts, sibilants, those live somewhere, and a scream is broadband with a peak. Worth checking where the peak sits before you build the whole library.



Corn

And that's a check he can actually run — plot the spectrum of a scream, plot where his model's confusions cluster, see if they overlap.



Herman

An afternoon of work, and it's the difference between augmentation and sabotage. Third limit. You were going to say the underfitting trap.



Corn

The one from earlier — it comes home for the solo practitioner. The published augmentation gains required bigger models and longer schedules. Daniel at home, fine-tuning a small model on his dictation, turns up the Ezra-noise, and his training loss climbs and his error rate gets worse. He'll think the technique failed. It didn't fail — he just didn't pay the entry fee. Harsh augmentation is a purchase, not a gift.



Herman

"My loss went up and my transcripts got worse" is the most predictable sentence in home machine learning.



Corn

Fourth, subtler — the fixed policy problem. Policy-SpecAugment, twenty twenty-two: instead of the hand-set mask parameters, they learn the policy, and on a hundred hours of LibriSpeech it beat standard SpecAugment by more than ten percent relative on test-clean. Their diagnosis was that a fixed policy brings relatively less data diversity in the low-resource setting — and low-resource is exactly Daniel's setting.



Corn

Which means when he copies the published parameters — the mask widths, the counts —



Herman

He's copying numbers tuned for someone else's model on someone else's data. The parameters are system-dependent. Copy the structure, not the numbers. He'd have to tune the mask sizes and the SNR sweep on his own setup anyway, which honestly is half the fun and all of the work.



Corn

There was one more finding in that low-resource paper that I want to pull forward — the superadditivity.



Herman

Oh, that's a lovely one. Combining the audio-modifying augmentations with the text-based ones — the ones that synthesize audio from transcripts — the improvement from the combination exceeded the sum of the individual improvements. Each method repaired the errors the other introduced. Text-based augmentation raised the frequency of those rare tone and palatalization cues by about thirty-seven, thirty-eight percent — patching exactly the holes masking had punched.



Corn

Two techniques, each covering the other's specific damage.



Herman

But — the superadditivity vanishes by a hundred hours. The complementarity is a low-resource phenomenon. Above a hundred hours, you're back to plain additive effects.



Corn

A hundred hours of what, though, speech? Because if Daniel's dictation corpus is well under that —



Herman

It almost certainly is, for a personal fine-tune. So he's in the window where combining methods pays extra. If he were doing this, the smart version isn't Foley alone — it's Foley plus masking plus maybe some synthetic text-driven audio, watched for whether the combination is still superadditive at his scale.



Corn

He'd enjoy that experiment more than the result. He always does. Which leaves the capture problem — the actual thing he flagged. You can't summon a scream.



Herman

No. And here's where I'd flip his framing slightly. He's been trying to capture moments on demand, like a photographer. Foley artists don't wait for moments — they stage them. You don't need Ezra mid-meltdown. You need a thing being thrown across a room. The sound of a small plastic block hitting a floor is a physical event you can reproduce any evening with a floor and a block.



Corn

You're suggesting Daniel spend his evenings throwing toys at the floor for science.



Herman

I'm suggesting he already lives with the source material and the acoustic space. The scream is the hard one, granted — you can't ethically stage a meltdown, and I wouldn't try. But screams are also the least maskable noise anyway; the realistic move there is capturing them opportunistically whenever they happen, which with a sixteen-month-old is a matter of when, not if. A recorder that runs when the dictation runs, kept out of the child's reach, quietly accumulating the real thing.



Corn

The microphone in the kitchen that Ezra already notices the voices coming from.



Herman

That one. The point is that "on demand" is the wrong requirement. The library builds itself opportunistically — you just have to be recording when it does.



Hilbert

You don't wait for it, Corn, it never comes.



Corn

...



Hilbert

The scream. You don't wait for it. I've had a recorder on my kitchen windowsill since March and the only thing worth keeping on it so far is Mr. Aboud's washing machine upstairs. Fourteen files now, one per cycle. Front loader. It does a high spin at the end that comes through the wall at a level I could recite for you, but I won't.



Herman

Is that what the windowsill recorder is for? It's been running since — that's your own noise library.



Hilbert

It's plugged into the timer that used to run the kettle. Herman's right about the microphone, that part he had correct, and I'll say the rest plainly: you don't get events by waiting. You record everything and cut what you need later, the way you always should have done. The machine does the waiting. Waiting is the machine's job.



Corn

The kettle timer.



Hilbert

It was doing nothing. Now it does something. The recordings are labelled by spin cycle — don't ask me why, they just are — and the neighbour knows, I told him the first week. He was pleased. He thinks someone should keep an eye on that machine, he's had two bearings replaced.



Corn

The label system. Why by spin cycle?



Hilbert

Consistency. Anyway, Corn, for all your long thinking on this segment, the practical part took me an afternoon and cost nothing. The recorder was forty-two shekels used, on the shelf at the place on Agrippas. It records a week on one set of cells.



Herman

Forty-two shekels and a windowsill, versus an eleven gigabyte download of six hours of somebody else's noise.



Hilbert

His neighbour's washing machine is better noise than MUSAN's. It has a bearing you can hear.



Corn

A machine doing the waiting while nobody watches. I'm not surprised it's his setup.



Herman

I nearly mentioned the deaf angle earlier and dropped it — MUSAN's noise portion, those six hours, was never auditioned for this job. Some of it is technical hum and babble from who knows where. Against that, a neighbour's actual washing machine in an actual wall — that's the Ivry mismatch argument in one appliance.



Corn

And it cost less than the tar file's electricity. Okay — the cutting-room floor. Mine today is the label smoothing footnote nobody reads: the SpecAugment paper mentions, almost in passing, that label smoothing combined with augmentation increased the proportion of unstable training runs on LibriSpeech. Two individually safe settings that destabilize each other — the fine print in this field is where the models go wrong.



Herman

Mine's the Switchboard policies, because they show the tuning instinct. Same paper, different corpus — the mask parameters drop from F27 to F15, the time mask budget changes, the fraction cap drops from one point zero to zero point two. Same authors, same technique, completely different numbers. If Google retuned it per corpus, Daniel retunes it per kitchen.



Corn

The forward thought, then. The personal Foley library is unexplored — the closest anyone's come is estimating your own room's impulse response, and that evidence says your own environment beats a borrowed one. The real test isn't the recording, it's the match: whether the noise you capture is the noise the model faces at inference. In Daniel's case it would be, almost by definition, which is exactly the advantage nobody publishing papers has.



Herman

And the saturation point cuts his way too — augmentation's benefit fades as data grows, and his data will never grow. The edge is largest precisely where he stands. Thanks as always to Hilbert Flumingtop at the desk, producing the show. If this was your kind of episode, go back for episode five, Fine-Tuning ASR For Maximal Usability; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts — if today's episode earned it, the review button is right there, and it takes ten seconds. We'll see you tomorrow.

SOURCES

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

1. **Park, Chan, Zhang, Chiu, Zoph, Cubuk, Le [primary]**
   SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, arXiv:1904.08779v3, submitted 2019-04-18, revised 2019-12-03, Proc. Interspeech 2019, 2613, 2617., retrieved 2026-11-21.
   https://arxiv.org/abs/1904.08779v3

2. **Snyder, Chen, Povey [primary]**
   MUSAN: A Music, Speech, and Noise Corpus, arXiv:1510.08484, 2015., retrieved 2026-11-21.
   https://arxiv.org/abs/1510.08484

3. **OpenSLR SLR17 [primary]**
   MUSAN download page, `musan.tar.gz` 11 GB, CC BY 4.0., retrieved 2026-11-21.
   https://www.openslr.org/17/

4. **torchaudio documentation, `docs/source/transforms.rst` (pytorch/audio repo; docs snapshot updated 2026-10-14) [primary]**
   `MyPipeline` SpecAugment example., retrieved 2026-11-21.
   https://github.com/pytorch/audio/blob/main/docs/source/transforms.rst

5. **Hugging Face dataset card `corypaik/musan`**
   subcorpus breakdown (music ~42h, speech ~60h, noise 929 files ≈6h)., retrieved 2026-11-21.
   https://huggingface.co/datasets/corypaik/musan

6. **audeering datasets**
   MUSAN card, version 1.0.0, CC-BY-4.0, 16 kHz mono., retrieved 2026-11-21.
   https://audeering.github.io/datasets/datasets/musan.html

7. **Ivry, Cohen, Berdugo**
   Evaluation of Deep-Learning-Based Voice Activity Detectors and Room Impulse Response Models in Reverberant Environments, arXiv:2106.13511v1, accepted ICASSP 2020., retrieved 2026-11-21.
   https://arxiv.org/abs/2106.13511v1

8. **Steinmetz, Ithapu, Calamia**
   Filtered Noise Shaping for Time Domain Room Impulse Response Estimation From Reverberant Speech, arXiv:2107.07503v1, WASPAA 2021., retrieved 2026-11-21.
   https://arxiv.org/abs/2107.07503v1

9. **Deppisch, Meyer-Kahlen, Amengual Garí**
   Blind Identification of Binaural Room Impulse Responses from Smart Glasses, arXiv:2403.19217v2, IEEE/ACM TASLP vol. 32, pp. 4052, 4065, 2024., retrieved 2026-11-21.
   https://arxiv.org/abs/2403.19217v2

10. **Li, Zhang, Zhuang et al.**
    Frame-level SpecAugment for Deep Convolutional Neural Networks in Hybrid ASR Systems, arXiv:2012.04094v1, SLT 2021., retrieved 2026-11-21.
    https://arxiv.org/abs/2012.04094v1

11. **Ibaraki, Chiang**
    Error Rate is Not Enough: Costs of Augmentation in Low-Resource ASR, arXiv:2610.13807v1, submitted 2026-10-09., retrieved 2026-11-21.
    https://arxiv.org/abs/2610.13807v1

12. **Li, Ma, Zhao et al.**
A Policy-based Approach to the SpecAugment Method for Low Resource E2E ASR, arXiv:2210.08520v1, APSIPA ASC 2022., retrieved 2026-11-21.
https://arxiv.org/abs/2210.08520v1

13. **Jin, Xie, Wang et al.**
Towards Automatic Data Augmentation for Disordered Speech Recognition, arXiv:2312.08641v1, ICASSP 2024., retrieved 2026-11-21.
https://arxiv.org/abs/2312.08641v1