Episode Book

EPISODE #6525

Why Speech Recognition Models Struggle With Accents

PUBLISHED	11 November 2026 at 14:52 UTC
DURATION	23:05
LISTEN	myweirdprompts.com/episode/speech-to-text-accents-bias-43640/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	25 cited by the research (listed at the end)
BOOK GENERATED	12 November 2026 at 12:58 UTC

SYNOPSIS

Speech-to-text models still benchmark against standard American and British English, treating other accents as margins of error. Longitudinal audits show newer systems often widen performance disparities, while decoder models like Whisper frequently insert hallucinated text rather than omit difficult words. Meanwhile, consumer interface options like locale dropdowns and prompt instructions fail to solve acoustic mismatches, offering marginal decorative priming instead of the structural changes seen in regional fine-tunes and specialized acoustic adapters.

TRANSCRIPT



Corn

There's no such thing as not having an accent. There's only having the one the machine was measured against.



Herman

That's the whole episode in one sentence, and it's also the trap. Because the interface everyone actually meets is a dropdown, and the dropdown is not where the work happens.



Corn

Daniel sent us a prompt about exactly this. How speech-to-text models have handled accents within English, whether having one costs you accuracy versus choosing, his words, just regular English, whether picking your own accent in a model that offers one actually biases it toward your speech or just makes you feel better, and whether the whole accent-selection feature fades away as models go multilingual and stop being told anything.



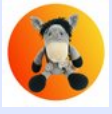
Herman

And there's a nice wrinkle in there. The phrase just regular English. He knows it's a fiction, that's why he put it that way.



Corn

Daniel's Irish. He's allowed one regular English per lifetime, and that was it.



Herman

So the structure of the answer is a little upside down. The dropdown looks like the feature, and everything we know says the real accent handling lives in the training data and the fine-tunes, a floor below where the user can reach. Let's take the gap first, then the dropdown, then the multilingual question.



Corn

Go.



Herman

Historically the default was one model, one standard, and everyone else measured as an error rate. There's a longitudinal audit accepted at SLT this year, Metzger and colleagues, sixteen systems released between twenty twenty-one and twenty twenty-five, eight accent datasets, and their framing is blunt. Standard ASR benchmarks, in their words, effectuate an Anglocentric language policy. Two accents function as the reference, Mainstream US English and Standard Southern British English, and every other variety of English is scored as deviation from them.



Corn

Scored as deviation. Meaning if you speak Irish English, your speech isn't a variety. It's a margin of error.



Herman

That's their argument, yes. They quote a line about how the absence of an official policy often serves only to reinforce the power and hegemony of prestige varieties and marginalize others.



Corn

My favourite kind of policy. The one that exists by denying it's a policy.



Herman

And the numbers back it up. Average gaps across the datasets run from nine percent absolute word error rate between Standard Southern British English and Irish English, up to twenty-six percent absolute between Mainstream US English and Vietnamese English on one of the ALLSTAR sections.



Corn

Twenty-six percent absolute. That's not a rounding problem. That's the difference between a transcript you'd send to your boss and one you'd send to your enemies.



Herman

Right, and it's not shrinking. The audit's most quotable finding is that minority-accent performance sometimes worsens by sixty-five percent absolute word error rate between generations of systems. Sixty-five. So you buy the new model, the one with the launch video, and your accent got worse.



Corn

Which is the exact opposite of the marketing story, where every release closes the gap because the model got bigger and more multilingual.



Herman

The audit tested that directly. Statistically significant improvement over time showed up in only two of their eight datasets, and in one of those, the four most recent models reversed the trend. Their phrasing is that the disparity is stubborn, not reliably declining as a function of time. And there's a companion finding I find strange. Audio compression predicts accent fairness more than model scale. Which says the codec you fed the audio through matters more than how many parameters are listening.



Corn

So the elephant in the server room is a compression artifact.



Herman

Whisper specifically. On the Edinburgh accents corpus, EdAcc, Whisper averages just under twenty percent word error rate. On clean US English read speech, two point seven. Sevenfold difference, same model, and the worst drops are for Indian, Jamaican and Nigerian speakers.



Corn

I want to stay on the failure mode for a second, because substitution isn't the interesting part. Getting a word wrong is something a human does too.



Herman

Agreed. The interesting part is what the decoder-style models do when they're lost. There's a paper from this year, Do LLM Decoders Listen Fairly, around forty-three thousand utterances, and on Indian-accented speech Whisper's insertion rate, text that appears in the transcript that nobody said, spikes to over nine percent at large-v3. And the spike is non-monotonic, which is the really damning part. Smaller model, fewer insertions. Bigger model, more confident invention.



Corn

The bigger model doesn't hear better. It hallucinates more fluently.



Herman

And it gets worse when you stress it. Inject silence into the audio and the accent bias amplifies by a factor of up to about four and a half, because the silence triggers demographic-selective hallucination. Under chunk masking they found repetition loops accounting for the overwhelming majority of insertions. So the model doesn't go quiet when it fails. It talks.



Corn

That's the part I'd want Daniel to hear, actually. The symptom isn't the machine missing your words. It's the machine adding words you never said, confidently, in a register that looks like a real transcript.



Herman

And second-language speakers get the worst of it across the board. There's an ICASSP paper this year that ran Whisper and Seamless-M4T across twenty-six accent groups and found, their words, large fluctuations in word error rate, indicating significant fairness gaps.



Corn

Twenty-six groups. Even the fairness analysis needs a spreadsheet.



Herman

There's one more number I'd sit with, though. An earlier audit of the Speech Accent Archive, twenty-seven hundred speakers, two hundred-plus first languages, found that the disparities correlate with the political alignment of the speaker's birth country with respect to US geopolitical power. Even controlling for the linguistic factors. So the error rate tracks geopolitics, not phonetics.



Corn

Which is either a profound finding about data or a very roundabout way of ranking allies.



Herman

Now the practical question hiding inside Daniel's prompt. How would you know you're on the wrong side of the gap? Because nobody ships you your error rate.



Corn

By comparison, I assume. You can't see your number. You can only see the difference.



Herman

That's the method the audit itself uses. MUSE as a control accent in every dataset, same task, and the gap is the finding. At home, the version is the same utterance under different locale settings, or the same sentence you and a MUSE-speaking friend both dictate. The symptoms are recognizable too. Proper nouns and names get mangled first. There's a whole project called AfriNames about ASR models butchering African names. Filler words and disfluencies get dropped. And in the bad cases, whole clauses of hallucinated insertion rather than omission.



Corn

So the diagnostic is: your name comes out wrong, your fillers vanish, and occasionally the machine invents a sentence for you.



Herman

That's a clean checklist, and I don't like how well it describes my own dictation.



Corn

Yours. You speak Hebrew with an accent nobody can place. Which, per this literature, makes you the least classifiable man in Jerusalem.



Herman

It's a gift. The model can't discriminate against my accent because it can't decide what it is.



Hilbert

Hilbert's Mid-Show Haiku. The dropdown. Fifteen Englishes offered. One of them is mine. The machine hears all of them wrong.



Corn

Fifteen, by the way, is a real count. Apple's dictation settings list fifteen English variants.



Herman

Six for Windows voice typing. Australia, Canada, India, New Zealand, UK, US. And that's the interface. Now the question Daniel actually asked. Does picking your own in that list do anything?



Corn

My instinct is that it does something, because it feels like it must. Every setting that exists exists for a reason.



Herman

That instinct is exactly what I'd want to check, because the direct evidence is thin to the point of vanishing. Nobody has published a controlled study of a consumer accent dropdown versus default English. The closest thing is a small Siri study from twenty twenty-four, thirty-six clips, thirteen accents, nine English locales, seven hundred transcripts, and the per-locale breakdown never surfaced in anything I could get at.



Corn

Seven hundred transcripts and the answer stayed inside the repository.



Herman

So you build the answer from the pieces around it. And the pieces point in an uncomfortable direction. When accent conditioning is built into the model, it works. There's an open project, whisper-accent, that freezes Whisper and feeds it learned accent embeddings through adaptive layer normalization. Fourteen percent word error rate on accented English for the small model, thirteen for medium, against about seventeen and a half for the same-sized vanilla models. Three or four points absolute from accent conditioning alone. That's real.



Corn

But that's the model being told, at the level of its normalization layers, what accent it's hearing. The dropdown is you telling it in a text box.



Herman

And there's the gap. Whisper has a prompt parameter, the thing people put transcribe in British English into, and OpenAI's own developer community describes it as not really accepting instructions, more for style, like punctuation. The prompt biases the decoder toward token sequences. It doesn't tell the model anything about acoustics.



Corn

Meaning the sentence transcribe in British English is doing the same job as the black cat played cheerfully with the pink elephant.



Herman

Functionally, yes. It's English text priming the decoder. The content is decorative. And it's worse than decorative, because the prompt eats tokens. There's a practical ceiling around two hundred and twenty-four tokens before you start truncating the audio context, so every word of your accent instruction is audio the model gets less of.



Corn

You're spending the machine's memory on a spell it can't read.



Herman

And there are documented snags. Users in Wales prompting with English context and getting their English speech transcribed into Welsh. Another one where English audio came out translated into Malay. The prompt can actively push the model into the wrong language, which is a spectacular own goal for a setting whose only job was to help your accent.



Corn

So the dropdown is at best a shrug and at worst a trapdoor.



Herman

At the same time, I don't want to overstate it, because prompting isn't nothing in general. Context-aware prompting work on Arabic reports around twenty percent relative error reduction on accented Modern Standard Arabic, and a Google paper last year found elicitor prompts improving Whisper by up to twenty-four percent relative. But that's contextual vocabulary biasing, guiding the decoder toward likely words. It's not accent selection, and the gains on dialectal Arabic dropped to under ten percent, which tells you where the ceiling is.



Corn

And at the other end, the intervention ladder. You said fine-tuning beats everything.



Herman

Consistently, in every paper I looked at. Regional adapters on a Qwen Omni model get six to ten percent word error rate on Indian, Indonesian and Latin American accented conversational English, and entity recall, the names and proper nouns, jumps from the mid-fifties to over eighty percent. Meanwhile the test-time trick of steering the model's internal representations toward an accent direction, there's a paper this year titled A Probe Shift Is Not a Fairness Fix, gets you under point-seven percentage points. Less than a point.



Corn

So the ladder is, fine-tune and get five points, prompt and get fractions of one. And the dropdown lives somewhere below the bottom rung.



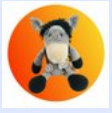
Herman

Below it, or beside it, doing a different job entirely. Some of those locale settings change which model runs. On Siri, English UK and English US historically weren't the same recognizer underneath, they were different downloads with different training. So the setting isn't fake there, it's a model swap, and the model swap is a real intervention. The placebo reading is when the setting feeds a prompt into a multilingual model that was never listening.



Corn

Which is increasingly the case. Everything new is multilingual and told nothing.



Herman

That's Daniel's last question, and it's where the terminology gets interesting, because he used the word alingual, and I went looking for it. No speech recognition source uses it. The only hit is a machine translation paper from two thousand and nine about alingual concepts, aligning parallel corpora. In ASR it doesn't exist as a technical term.



Corn

So Daniel coined it, or imported it from the commentary side.



Herman

From the LLM discourse, I'd guess. It's a good coinage, it just isn't a category anyone in speech works with. What exists is the real thing underneath it. Whisper, Seamless-M4T, the models in the audit's table going up to one literally named Omnilingual in twenty twenty-five. All of them infer language and variety from the audio rather than being told. Their whole architecture is agnostic to your dropdown.



Corn

So the structural argument writes itself. If the model auto-detects, the dropdown is redundant at best.



Herman

Redundant, or a mis-specification hazard. Here's the counter-intuitive part, though. Accent identification isn't being removed from these systems. It's being built in, internally. There's a design from twenty twenty-one, Dyn-ASR, that runs language ID and accent ID on-device to route audio to accent-specific models, using under a twelfth of the memory of the alternatives. Newer work trains an accent classifier and then masks the spectrogram regions driving its prediction, so the recognizer becomes accent-invariant by literally hiding accent from itself. And there's a paper this year generating speaker embeddings from natural-language prompts. A thirties male speaker with an Indian accent, in as text, out comes a synthetic profile.



Corn

Wait. The accent prompt doesn't work when the user types it, and works beautifully when it's typed into the embedding generator. Same sentence, different plumbing.



Herman

Same words, completely different mechanism. One biases token sequences. The other steers the model's speaker representation. That's the whole tragedy of the dropdown in one comparison.



Corn

So the honest forecast is the labels go away and the conditioning stays.



Herman

The labels migrate from the settings screen to the pipeline. The user-facing feature declines, yes, I'd bet on that, multilingual models make it redundant and vendors hate maintaining locale lists nobody understands. But accent conditioning becomes automatic, invisible. And the catch is that everything now depends on the auto-detected accent being right, and the hallucination evidence says detecting variety is precisely where these models fail most visibly.



Corn

The machine that can't hear a Vietnamese English speaker correctly is now also the machine deciding, silently, what accent it's hearing.



Herman

At least the dropdown told you what it thought. Badly, but in writing.



Corn

Herman, one more thing you said earlier. You called the dropdown the user-facing accent choice. But there's no setting for a foreign accent in English at all, is there. Fifteen variants, and they're all regional. Indian English is on Apple's list because India is a market. Vietnamese English isn't on anyone's list.



Herman

No, and that's a sharp asymmetry. The dropdown offers the accents that map onto national markets. The accents that suffer most, second-language varieties, the ones in the twenty-six percent gap, have no row in the settings at all. The feature doesn't fail them. It declines to acknowledge them.



Corn

Twenty-six percent, and the menu has nothing for them. That's the part I'd call a policy.



Herman

Herman's... no, go on.



Corn

I was just going to ask what you'd tell Daniel to actually do. Irish English, the nine percent gap end of the range, on a model that mostly works for him.



Herman

Honestly? The same utterance under two settings, compared. It's a two-minute experiment and it's the only diagnostic anyone has. And proper nouns as the test case, because that's where the gap shows first. If the names survive, stay. If the machine starts writing sentences for you, the setting isn't your problem, the model is.



Corn

Don't ask whether the setting works. Ask whether your name comes out spelled right.



Hilbert

Corn, how many Englishes do you suppose the council's phone system offers? I'll tell you. None. You get Hebrew, Arabic, Russian, and a thing they call English.



Herman

The recordings?



Hilbert

The recordings are fine. The problem is the other end. I called the municipal records office three weeks ago about a rates letter, and they've moved to a voice system that types what you say and files it. The operator reads your words back off a screen now. Doesn't listen anymore. The machine listens.



Corn

And the machine doesn't care for your English.



Hilbert

The machine has decided I am difficult. The operator is a patient man named Tawil. He reads back what the machine wrote and I correct him and he corrects the form and the machine overwrites the correction on save. We have been through this four times. My surname is Flumingtop. The machine is committed to Fleming-Topp with a hyphen it invented. Tawil says the hyphen is not personal.



Corn

He's defending the machine's honour.



Hilbert

He's been there nineteen years, he says, and it's the first system that types for him, and it's saved him the writing, so he keeps it. The letter is still wrong. The rates stand at forty-one shekels a month and I am not paying a surcharge on a hyphen.



Herman

Did you try spelling the surname to it?



Hilbert

I did that in week two. You spell to the machine and it writes the spelling as a word. I got a letter addressed to Eff Ell You Em. That one I kept. The original rates letter too, the paper one from before, when a person typed it and got the name right. The paper is better



Hilbert (continued)

than the system they replaced it with.



Corn

That's the part I believe most. The old letter's right, the new letter's wrong, and you're paying the difference.



Hilbert

Tawil does his best. He phones me back on Fridays to say it's still not fixed. That's more than the machine does.



Corn

Forty-one shekels a month and a man who phones you back on Fridays. The paper's right and the system's wrong, and only one of them knows it.



Herman

The rates letter's actually a good small picture of the whole thing. The machine writes what it already believes the name should be. The correction doesn't survive contact with the save button.



Corn

The man who knows it's wrong keeps the system anyway, because it saved him the writing. That's the adoption problem in one office.



Herman

If I take one thing from today, it's the size of the intervention ladder. Accent conditioning inside the model buys you three, four, five points. Steering the representations buys you under one. And the setting on your phone buys you a feeling. The distance between what works and what's offered is the whole story.



Corn

Mine is the sixty-five percent. A new generation of model made a minority accent worse by that much. Anyone waiting for the upgrade to fix their accent is waiting on a trend that isn't there.



Herman

The open question I'd leave with Daniel. When the dropdown disappears and the accent conditioning goes internal, who checks that the auto-detected accent was right? The model that hallucinates on the accents it mishears is the same model now deciding, silently, what it's hearing. Somebody has to run the comparison, and at the moment that somebody is you, with two recordings and a Saturday afternoon.



Corn

If you take one thing from this, make it the checklist. Your name, your fillers, your invented clauses. That's how you know.



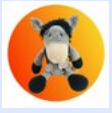
Herman

If you want to go deeper, we've talked about ASR before, how it got good, what happens to a fine-tune when a voice ages, and what to do after you've fine-tuned one. Those are in the archive.



Corn

Producer credit to Hilbert Flumingtop, who kept the letter.



Herman

If this was your kind of episode, go back for episode ten, How ASR Went From Frustration To … Whisper Magic; episode five, Fine-Tuning ASR For Maximal Usability; and episode seven, Building Custom ASR Tools. This has been My Weird Prompts, the human-AI collaboration podcast. Find us at my weird prompts dot com, and leave a review if the mood strikes.



Corn

We'll be back soon.



Herman

See you tomorrow.

SOURCES

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

1. **Metzger et al., Connecting the Dots: Enduring Performance Gaps in ASR Systems Between Standard and Minority English Varieties, arXiv:2610.17761v1, submitted 2026-10-13, accepted IEEE SLT 2026. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2610.17761

2. **Performance Disparities Between Accents in Automatic Speech Recognition, arXiv:2208.01157v1. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2208.01157v1

3. **Ginjala, Fosler-Lussier, Myers et al., Do LLM Decoders Listen Fairly?, arXiv:2604.21276v1, submitted 2026-04-23. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2604.21276v1

4. **Bourrel, Issam, Spanakis, A Probe Shift Is Not a Fairness Fix, arXiv:2609.18533v1, submitted 2026-09-16, IMPACT-SPEECH 2026. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2609.18533v1

5. **Swain, Maji, Mishra et al., Towards Fair ASR For Second Language Speakers Using Fairness Prompted Finetuning, arXiv:2510.18374v2, revised 2026-01-25, ICASSP 2026. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2510.18374v2

6. **Husain, Pandey, Singh, Beyond WER: Entity and Disfluency Recall in Accented Conversational ASR, arXiv:2609.20828v1, submitted 2026-07-22, Interspeech 2026. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2609.20828v1

7. **The Edinburgh International Accents of English Corpus (EdAcc). Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2303.18110

8. **MLS-trained ASR accent evaluation. Retrieved 2026-11-11. [primary]**

   http://arxiv.org/pdf/2105.05041

9. **Patel & Scharenborg, Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition, arXiv:2307.02009v1, 2023-07-05. Retrieved 2026-11-11. [primary]**

   https://arxiv.org/abs/2307.02009v1

10. **Olatunji et al., AfriNames: Most ASR models 'butcher' African Names, arXiv:2306.00253v2, Interspeech 2023. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2306.00253v2

11. **Parcollet et al., Loquacious Set, arXiv:2505.21578v1, Interspeech 2025. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2505.21578v1

12. **Ghangam, Whitenack, Nemecek, Dyn-ASR: Compact, Multilingual Speech Recognition via Spoken Language and Accent Identification, arXiv:2108.02034v1, IEEE WF-IOT 2021. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2108.02034v1

13. **Sameti et al., Accent-Invariant Automatic Speech Recognition via Saliency-Driven Spectrogram Masking, arXiv:2510.09528v1, 2025-10-10. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2510.09528v1

14. **Thebaud et al., ProPS: Prompted Profile Synthesis, arXiv:2607.05276v2, SLT 2026. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2607.05276v2

15. **Jin, Srirag, Joshi, PAREDA: A Multi-Accent Speech Dataset, arXiv:2605.17860v1, SPEAKABLE 2026 @ LREC 2026. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2605.17860v1

16. **Context-aware decoding for Arabic ASR. Retrieved 2026-11-11. [primary]**
https://arxiv.org/pdf/2511.18774

17. **Whisper Accent [primary]**
Accent-Aware English Speech Recognition. Retrieved 2026-11-11.
https://github.com/mavleo96/whisper-accent

18. **Cheng (Google), Context-Aware Speech Recognition Using Prompts for Language Learners, Interspeech 2024. Retrieved 2026-11-11. [primary]**
https://www.isca-archive.org/interspeech_2024/cheng24b_interspeech.pdf

19. **Assessing Smartphone Speech Recognition across Diverse English Accents: A Preliminary Study (Siri, 36 clips, 9 locales, 702 transcripts). Retrieved 2026-11-11. [primary]**
https://iris.cnr.it/handle/20.500.14243/529673

20. **Adapting Whisper for Regional Dialects: Enhancing Public Services for Vulnerable Populations in the United Kingdom, VarDial 2025. Retrieved 2026-11-11. [primary]**
https://aclanthology.org/2025.vardial-1.4/

21. **Wee et al., Adapting Automatic Speech Recognition for Accented Air Traffic Control Communications, arXiv:2502.20311v1, 2025-02-27. Retrieved 2026-11-11. [primary]**
https://arxiv.org/abs/2502.20311v1

22. **OpenAI Developer Community thread, posted 2024-07-12. Retrieved 2026-11-11.**
https://community.openai.com/t/whisper-transcription-translates-to-random-language-malay/863452

23. **OpenAI Developer Community thread on Whisper prompts and Welsh. Retrieved 2026-11-11.**
https://community.openai.com/t/i-live-in-wales-but-speak-english-chatgpt-is-confused/807620?page=2

24. **DEV Community, Using LLM for Speech Recognition. Retrieved 2026-11-11.**
https://dev.to/shashank_ms_6a35baa4be138/using-llm-for-speech-recognition-a-comprehensive-guide-1aob

25. **Microsoft Q&A on Whisper source-language hallucination. Retrieved 2026-11-11.**
https://learn.microsoft.com/en-us/answers/questions/2182774/azure-openai-whisper-hallucinates-source-audio-lan