Episode Book

EPISODE #6663

Blind-Testing 26 Hebrew Text-to-Speech Models

PUBLISHED	18 November 2026 at 21:13 UTC
DURATION	38:17
LISTEN	myweirdprompts.com/episode/hebrew-tts-blind-test-46965/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	omnivoice
PIPELINE	V5.35
BOOK GENERATED	18 November 2026 at 22:10 UTC

SYNOPSIS

Twenty-six text-to-speech models went into a blind bake-off across sixty runs to find a reliable Hebrew pronunciation coach, and twenty-three runs failed at screening. Testing sentences both with and without nikud vowel points reveals how much models guess, from Kokoro Hebrew reading 'book' as 'barber' to Zonos 3 Mini falling back to English. The episode explores why Hebrew's missing vowels, syllable stress, and scarce studio data like SASPEECH make natural speech synthesis so difficult.

TRANSCRIPT



Corn

Twenty-six text to speech models, sixty runs, one man with a headset and a lot of opinions about the letter R. Daniel spent today running a blind bake-off of Hebrew voices, and he wants us to walk through what he found, clip by clip, because the answers are a lot stranger than "this one is good, this one is bad."



Herman

And the reason he did it is practical, which I like. He's building Hebrish, a speech recognition project for people who flip between English and Hebrew mid-sentence, and he's fine-tuning his own Hebrew speech to text model on his own recordings. The recorder app has a Hebrew project where he reads sentences aloud off the page. Daniel is not a native Hebrew speaker, so the page has nikud on it, the little vowel points, and next to each sentence there's a Listen button that plays a native voice first, so he knows what he's aiming at.



Corn

A pronunciation coach that's also a text to speech model. So if the coach is wrong, he trains himself wrong.



Herman

Worse, he trains the speech recognizer on himself being wrong. The vowels come from DictaBERT, Dicta's diacritizer, which he rates highly and I'd agree. But he checked every sentence word by word anyway, and he corrected twelve words across eight of the twenty test sentences. The classic one is the word spelled samekh, peh, resh. Read it sefer and it's a book, read it sapar and it's a barber. Same three letters.



Corn

Same letters, different haircut.



Herman

The other one is hergel against ha-regel. Hergel is a habit, ha-regel is "the leg." The model picked the leg for a sentence about habit. If you read that aloud with confidence, you've learned a sentence that means something else.



Corn

What was the old coach voice like?



Herman

It was Phonikud feeding a Piper voice, pre-rendered sentence by sentence. Daniel's complaints were that it was flat, that the newer sentence sets had no audio at all, and that it stressed words on the last syllable. Sefer comes out as sef-ER, yeled, a boy, comes out as yel-ED. Nikud marks vowels but it doesn't mark stress, so a system that only has the points to go on has to guess where the weight falls, and it guessed wrong. The sentence he used most in the bake-off is the first one, "Ani hayiti ose et ze shuv." I would do it again. Listen to the very end of it, where the last word goes somewhere it shouldn't. Roll the old reference voice.



Corn

It goes up. It asks a question it was never asked.



Herman

Daniel's note on that was that "I would do it again" comes out with fairly flat prosody, and then that strange lift at the end. A man who isn't sure he would do it again.



Corn

Which is a fair review of the sentence, if you've been reading Hebrew with a model in your ear.



Herman

So he went looking for a replacement, and he wanted it measured properly, because he'd rather not rely on a feeling. Twenty sentences, plus one longer paragraph he calls paragraph one for the head to head rounds. Ten of the sentences come from Tatoeba, two at each of five difficulty grades, and ten he wrote himself, each aimed at one specific trap. Segolate stress, the sefer and sapar pair, times of day spelled out in words, a foreign surname, in this case "My name is Daniel Rosehill and I live in Jerusalem." Technical talk, so "I'm training a speech recognition model on my own recordings." Forklifts and pallets. English loanwords, "I sent you an email with the link to the file on the Drive." And one long sentence with a lot of commas.



Corn

He tested the forklift sentence on a speech model.



Herman

The warehouse people deserve a voice too. And every sentence exists twice, with nikud and without it. That gives you the thing I think is the intellectual centre of the whole exercise, which is nikud inference. Anyone reading unpointed Hebrew has to infer the vowels, a person or a model. So if a model accepts both forms, you run it twice, plain and pointed, and the gap between the two runs is how much that model is guessing.



Corn

And if the gap is huge, the model was never reading Hebrew, it was doing a very confident impression of it.



Herman

Kokoro Hebrew is the neat example. Given plain text it read the word for "the book" as ha-sefer, correct. Given the nikud, it read ha-sapar, "the barber." Its phonemes differ between the two inputs in sixteen of the twenty sentences. So handing the model more information made it worse on that one.



Corn

Which is a bit like giving directions to someone and they get more lost.



Herman

Then the judging. Screening first. Every clip is a pass or a fail, blind, model names hidden, order shuffled. A fail gets exactly one reason, mispronunciation, wrong stress, a glitch, truncated words, wrong voice or language, pacing, or accent. A run is out at its first fail. The survivors go into an arena, head to head on the paragraph, blind again, and a Bradley-Terry fit turns the results into a ranking.



Corn

Judging a voice when the vowels are all over the place is its own problem.

L

Larry

Are you tired of your text to speech models arguing over nikud like it's a family wedding? Introducing VowelVarnish, the only blind bake-off spray that judges twenty six voices at once and crowns a winner before the vowels even show up. Just one spritz and every Hebrew sentence comes out with perfect cantillation, a slight Sephardi lilt, and a certificate of authenticity signed by a man named Gemini Flash who does not exist.

L

Larry

VowelVarnish is not responsible for sudden fluency in Aramaic, spontaneous aliyah, or any vowel you may have previously considered optional. Results vary by cantor, altitude, and whether your sloth is watching. Side effects include excessive diacritics, a lingering smell of burnt phonemes, and the unshakable feeling that you are being judged blind. Do not use VowelVarnish while operating a donkey. Ask your producer if VowelVarnish is right for you, and remember, no refunds, no vowels, no problem.



Corn

I am watching, for the record.



Herman

The judging standard is what makes it a hard test. Daniel's rule is neutral modern Israeli pronunciation. If you hear a trilled R, a pharyngeal het or ayin, an Ashkenazi reading of the vowels, that's a fail, even though each of those is real Hebrew somewhere, from some community, in some decade.



Corn

So a voice can be perfectly good Hebrew and still fail. That's a harsh grader.



Herman

He's the customer. The Listen button has to sound like the person he'd meet at the supermarket in Jerusalem, not a cantor and not a newsreader from nineteen sixty-eight. And most of the annotation is done by voice. He presses R on a clip, says what he heard, presses R again, and the note uploads with the clip. Later Gemini transcribes it and structures it. The app is open source, it's his TTS pronunciation annotation starter. Four of those voice notes are going to play in this episode, mid-thought, blind, so you can hear him at the desk.



Corn

You've covered Hebrew speech models here before. The ears, the mouth, the brain, and the little models that sit behind the show's voice. I'd rather not rerun that.



Herman

We'd be repeating ourselves, yes. In those episodes we looked at the arena standings, at Phonikud's approach and at the nikud models. Today's question is different, because the field is much bigger and nobody had put it in a room and listened blind. So, first, why is Hebrew a hard case at all? Plenty of languages lack vowels in the writing.



Corn

Does that include English? I've seen people write "txt" and survive.



Herman

English text messages are a choice. Hebrew is an abjad. Everyday printed Hebrew leaves out most vowels as a rule, and readers fill them in from context so fast they don't notice. For a human that's elegant. For a model it's a pile of homographs. Then the fix, nikud, is a patch with holes. It leaves out stress. It's mostly built on formal grammar rather than how people speak. Diacritized text is scarce, so the diacritizers themselves are trained on a thin diet.



Corn

Two problems stacked, then. Which vowel, and which syllable gets the weight.



Herman

Three, if you count the prefixes. The word for "the", "and", "to", "in", "that" all glue onto the front of the next word, so the boundary the model needs isn't on the page. And there's a spoken versus written gap. Acronyms get read as words, army names, institutions, and loanwords are irregular. Daniel's "email" sentence is deliberately a trap for that.



Corn

Then there's the data. I assume English has a mountain and Hebrew has a molehill.



Herman

The cleanest studio corpus for Hebrew TTS is called SASPEECH, a single speaker, a podcast host, about thirty hours, recorded in a studio, with a custom licence that doesn't allow commercial use. English has a famous single speaker set called LJSpeech at about twenty-four hours. So Hebrew's best clean corpus is about the size of what English researchers use as a classroom exercise. Then there are enormous Hebrew sets, thousands of hours, but they're built for speech recognition: uncontrolled audio, a thousand speakers, transcripts produced by another model. Fine for teaching a machine to listen. Not what you'd choose to teach it to talk.



Corn

Which explains something Daniel found, I think.



Herman

It does. No model in his field was trained from scratch on Hebrew audio alone. Every Hebrew-specific model is a fine-tune of an English or a multilingual model, or multilingual from the start. Kokoro Hebrew is a fine-tune of Kokoro on that single speaker, so it inherits the voice and the licence. Piper "Noa" started from an English LJSpeech voice and was pushed toward Hebrew on synthetic Hebrew audio. F five Hebrew, a hundred and fifty-eight hours of fine-tuning.



Corn

Give me a case where the stress actually changes what you're saying. Not a vibe, an actual word.



Herman

Tahini. In one of the Phonikud papers the example is a pair spelled identically, the sesame paste against the word for grinding, and the only difference is where the weight falls. The beer and capital city pair, bira, is another. Then gender. Hebrew verbs and adjectives are gendered, so "you wrote" is a different word for a man and a woman, and a model reading unmarked text has to decide who's being spoken to. One Tel Aviv vendor, Deepdub, says its model picks the right form instead of defaulting to masculine. That's a claim from a company, and I haven't seen anyone test it independently.



Corn

So the sentence you're reading can quietly be about a different person.



Herman

And nothing on the page tells you.



Corn

A premise worth checking, then. People say few models support Hebrew.



Herman

Backwards, really. Lots of models list it. Zonos lists forty-three languages, Chatterbox multilingual twenty-three, VoxCPM two thirty, Fish Audio eighty-plus, OmniVoice over six hundred. The claim to make is that very few do it well. Support is a line in a README. Daniel's bake-off is about the gap between the line and the sound.



Corn

Six hundred languages on the box, and the box is the best part. I'd like a refund on the other five hundred and ninety-nine.



Herman

Fair, though the six hundred one gets only one fault, which we'll come to. He ran about twenty-six models, ten with open weights and the rest commercial APIs. Some came as several runs, plain text and nikud, female and male voices, which is how twenty-six models turn into sixty runs. Twenty-three of those runs were gone at screening. The whole commercial set cost under a dollar to generate, and the entire bake-off a few dollars of cloud compute. Which is the other thing I find striking. You can do this properly for the price of a coffee. Start with the one that doesn't speak Hebrew at all. Zonos three Mini, from Zyphra. Same sentence, "Ani hayiti ose et ze shuv." Listen for what language it thinks it's speaking. Roll the first mystery language.



Corn

That's English. An American voice reading a transliteration of something it's never seen.



Herman

It had no Hebrew voice, so it fell back to its American English release voice. Eight seconds of confident nonsense. Daniel's note on it was recorded blind, he didn't know which model it was. Roll Daniel's first verdict.



Corn

"This isn't even Hebrew. It came out as a non-intelligible language." He didn't even call it bad Hebrew.



Herman

Which is the right call. A hard fail, and the right fail category too, wrong voice or language. Forty-three languages on the label, and the one he needed wasn't one of the voices.



Corn

Next victim.



Herman

Chatterbox Multilingual, from Resemble AI, plain text. A footnote here. Chatterbox has a built-in Dicta nikud step, and in the current release it's broken, so this was the model's raw guess at the unpointed letters. And remember that Chatterbox is the model our own production pipeline has used for English. So no jokes about Resemble.



Corn

I'll make jokes about the output instead.



Herman

Listen for the first word. Roll the second mystery language.



Corn

It started with "ani." I'll give it that. Then it left the building.



Herman

Daniel's words were that it started with an "ani" and then it was not intelligible as anything, in any language he knows. A model that said it had Hebrew support and fails catastrophically. The footnote is that given the nikud it passed screening, but in the arena he called it robotic, staccato, clipped. You'll hear that comparison in a few minutes.



Corn

So the same model is a catastrophe or a metronome depending on what you hand it.



Herman

That's nikud inference showing up again. Next, Fish Audio S two Pro, given nikud. Here the Hebrew is closer to Hebrew. The question is what it sounds like. Daniel's first impression is in the name of the clip. Roll the Farsi clip.



Corn

I hear it. There's a lilt in there. Hebrew and Farsi are different language families, but I can hear why he reached for it.



Herman

He said every language has its vocal range signature, and this one sat in the wrong one. Then he went through the sentence against the nikud, and found two actual mispronunciations. Roll Daniel's Farsi verdict.



Corn

Osheh instead of oseh. And "shoob" instead of "shuv." So it wasn't a vibe failure, he found the exact words.



Herman

Which is what the blind process is for. It forces you to write down why. Fish was awkward to run, too. Self-hosting it took about thirteen minutes per sentence on a mid-range GPU, so he went through Fish's API, and the WAV headers it returned made his analysis tool try to allocate sixteen gigabytes of memory.



Corn

Sixteen gigabytes for a sentence about doing something again.



Herman

Next one. VoxCPM two, from OpenBMB, a different sentence, "It's not difficult if you study." One and a half seconds. Daniel's note is going to tell you the pronunciation was right, and the problem was elsewhere. Listen to the quality of the sound rather than the words. Roll the telephone clip.



Corn

Ah. That's a man on a payphone with a bad line.



Herman

Narrowband, tinny, nothing below the middle of the range. Let's hear how he put it. Roll Daniel's telephone verdict.



Corn

"A slightly angry man shouting down a bad telephone connection." The word "slightly" is doing real work in that sentence.



Herman

It's the kind of failure a word error rate would never catch. Every phoneme is right, and you'd still never put it in a product.



Corn

What happened to the others? You said twenty-three runs out.



Herman

The rest go by quickly. The old reference voice has a cousin, Phonikud with its own diacritizer feeding Piper, and Daniel said it had a French feeling, like a French Hebrew speaker, with awkward pacing. F five Hebrew version two, the one with the hundred and fifty-eight hours: plain, it started with the right word and then words went missing, truncated. Given the nikud the pronunciation was correct but very noisy, and its raw output was almost silent, around minus fifty LUFS.



Corn

Which makes a convenient excuse for a bad performance. Nobody can hear it.



Herman

Then OmniVoice, the one with over six hundred languages, failed on a pause after "ani" that wasn't in the text. What I liked is Daniel's reflection afterwards. He said there's an element of annotation that's subjective, and that's why he likes doing it, because if the pattern repeats, you get random pauses. One pause is a taste. Ten is a defect.



Corn

So the blind test finds the taste first and the defect later.



Herman

Yes, and he gives himself a way out of his own bias by running more than one sentence. Blue V three, the newest open model, released on the fourteenth of November, rose really, really up at the end of one sentence, "ze lo kashe im ata lomed", it isn't hard if you study. The same rising Phonikud and Piper did with the old voice, only louder.



Corn

Everyone's asking a question.



Herman

Kokoro Hebrew had the two faces I mentioned. Given nikud, wrong stress. Given plain text, it passed screening and then lost the arena for being very rushed. So there's no setting where it was both safe and pleasant. And Chatterbox, I said, passed with nikud and lost on sound. Do you see the pattern in that?



Corn

Every model has a way of being fine on paper. The plain version tells you what it does alone, the nikud version tells you what it does when you hold its hand.



Herman

ElevenLabs version three had a trilled R in "tishbor et habeitsim", you'll break the eggs. Daniel said it sounded like a multilingual model that advertised Hebrew and hadn't had a lot of training on it. It isn't wrong Hebrew, in fact, a trilled R exists. It's just not neutral Israeli, so it fails his rule.



Corn

A very good Hebrew voice, for the nineteen-thirties.



Herman

OpenAI's GPT four o mini TTS had awkward pacing, and with the nikud, Daniel called it kind of comical, the pacing all over the place. The commercial voices from Google Cloud, Inworld and MiniMax were out on mispronunciation, MiniMax also on wrong intonation.



Corn

And the one you've been dreading. The eventual winner failed too, didn't it?



Herman

It did, and I think it's the most useful clip in the whole bake-off, even though we can't play it. Gemini three point one Flash TTS, plain text, no nikud. Daniel caught it saying "hamatana hazot i lechaver sheli", with an "i" where the sentence has "hi", the word for "she." A single letter's worth of sound. He called it a tiny one. Given nikud, it never failed once.



Corn

So the best model in the room still needs the vowels written in.



Herman

One vowel, and the nikud makes it go away. Now, the middle of the field. Piper Hebrew, the voice called Noa, given nikud. It's the same Piper family as the old coach voice, but a different voice. It passed six screening sentences and won one arena match. It's a



Herman (continued)

paragraph, so a good test of pacing. Listen for the speed and for how little the sentences breathe. Roll the older generation paragraph.



Corn

It sounds like a very efficient person reading out a delivery note.



Herman

Daniel's note was significantly more robotic, sounds like a preceding generation of TTS, too fast, flat prosody, minimal pauses at punctuation. And the reason you don't throw it away is the price. About a hundredth of a cent per thousand characters on a CPU. The cloud voices run between one cent and eight cents for the same thousand characters.



Corn

Does anybody buy a voice that costs nothing and sounds like a delivery note?



Herman

Probably for something where the voice isn't the point. Phone menus, alerts. Not a Listen button for a learner, where the whole thing is feeling for where the sentence rises and falls. Let me ask you, Corn, as someone who naps through most of what he hears. How much does prosody matter to you, as a listener, if the words are right?



Corn

More than words. I'll forgive a mispronounced noun. I won't forgive a voice that sounds like it's reading me my rights. If the melody's wrong, I don't believe it's a person, and I stop listening.



Herman

Good, because that's precisely where the winner separates from the pack. The female bracket first, because every voice in the first arena round was female. Gemini three point one Flash TTS, voice Kore, given the nikud, on the paragraph. Roll the winning paragraph.



Corn

That's a person. It breathes, in places a person would breathe. The pace moves.



Herman

Unbeaten. Six wins, over Piper, xAI, Gemini two point five Pro, gpt-audio one point five, Gemini three point eight Live and ElevenLabs version four, and six ties. It costs about three cents per thousand characters, and it now runs the Listen button in the Hebrish recorder. Now Daniel's note, recorded in the arena, blind, comparing two clips he thinks are A and B. B is Chatterbox. Roll Daniel's verdict on the winner.



Corn

"Maybe some of the best Hebrew synthesis I've heard." And B is "robotic, staccato, clipped." The A wins in his mind before the clip even finishes.



Herman

Which is another way of saying that the winner is the thing everyone else is being compared against. And here's the part he did next. All of those first arena voices were female, and he realised a male voice against a female voice mostly measures the voices, not the models. So the evening of the same day he split the arena by gender and reran each unbeaten model with a male voice. Gemini as Charon, ElevenLabs as Brian, OpenAI as Cedar.



Corn

A man's bracket, a woman's bracket. And who's in the male final?



Herman

Nine of twelve male runs passed screening. The first match was won by Gemini three point eight Flash TTS in the Charon voice, on plain text, against Gemini three point eight Live. Roll the male paragraph.



Corn

Slower. A little warmer. It sounds like it's listening to itself.



Herman

Daniel's words were good, very listenable, a bit less robotic, more prosody and variation, a bit more personality, and he added that personality is the hardest thing to get in TTS.



Corn

Hold on. Before the family tree. Why does the blind part matter so much? He knows which model he's listening to for a second, then forgets.



Herman

Because he'd go in with opinions. He has used these brands for years, and when he sees a famous name on a clip, he hears it differently. Hiding the names turns "I like ElevenLabs" into "I like this clip", and the two come apart more often than you'd expect. It's also why the first arena was lopsided. With the names hidden and every voice female, he couldn't tell how much of his preference was the model and how much was just a timbre he happened to like. Splitting by gender was his fix, and he did it the same evening, which I think is the right instinct. When a test turns out to be measuring something other than what you meant, you change the test.



Corn

And you've watched him do that? Change the test halfway?



Herman

I can only say what the notes show. The male bracket opened the evening of the same day, with Gemini as Charon, ElevenLabs as Brian and OpenAI as Cedar, and nine of the twelve male runs passed screening, so the male voices were in better shape at the gate than the first female round.



Corn

Okay, so the obvious question. Female winner is three point one Flash, male winner is three point eight Flash. Are we being told it's the same model?



Herman

In effect, the same family. Google's Gemini Flash TTS line, different generations on each side, and three point one Flash in a male voice hasn't been screened yet. What I'd stress is that this agrees with a completely independent method. The ivrit ai Hebrew arena, where the crowd votes blind, also has Gemini on top, and the places behind it are proprietary too: Deepdub, ElevenLabs version three, Soniox. The one open entry, Blue version two, ranked eighth of twelve. Two methods, one winner.



Corn

Neither is perfect. A crowd is noisy, and Daniel is one man with a headset.



Herman

A crowd and a single careful listener disagreeing on the winner would be a story. Agreeing is the quiet kind. And Daniel didn't rely on his ears alone. Beside every verdict he logged loudness, clipping, pace, pitch, brightness and clarity per run, which is how the almost silent F five output and the telephone voice got numbers as well as adjectives.



Corn

The conversational models are in that list too, I noticed. The ones built to chat.



Herman

Gemini three point eight Live and the two point five native audio model, plus OpenAI's gpt-audio one point five. They're conversation models that he instructed to read a text verbatim, and the Live one stayed unbeaten in the female bracket. Which is a surprise if you think of text to speech as a separate product from talking. The line between them is blurring.



Corn

Why Gemini, though? Google isn't known for Hebrew.



Herman

Now I'm inferring, and I'll say so. Nobody has published a technical paper on how the Gemini voices are built, or why they're good at Hebrew. What we do know is the description Google gives. Gemini's speech models are native audio language models. The model reasons and generates speech natively in audio, rather than text going into a separate acoustic model. My guess, and it's only a guess, is that this helps a thin-data language. A language model has read vastly more Hebrew text than any TTS corpus contains, so it can use context to decide a homograph. It can tell the barber from the book because it understood the sentence, not because a dictionary told it.



Corn

That would also explain why it still slips on "hi" and "i". Context only gets you so far.



Herman

Yes, and it's a point in favour of the guess, because a model that understood the sentence perfectly wouldn't need the nikud, and it still does. Plain text, a vowel slips. Given the points, flawless. So even the best model is, in part, relying on the diacritizer, which brings the whole bake-off back round to where Daniel started. Dicta's check, twelve corrected words in eight sentences, wasn't a nuisance. It was the foundation.



Corn

And you said the cheap one is at par.



Herman

Gemini three point eight Flash Lite, tied the leader twice, at one cent per thousand characters, a third of the price. Daniel's words, a very good voice, a bit flatter. For a lot of uses that's the answer. It means that quality isn't the cost question any more. A Listen button for every sentence in an app costs almost nothing.



Corn

A year ago I'd have said a good Hebrew voice on every sentence of an app was a budget line with its own meeting.



Herman

Now it's rounding error. A sentence runs to maybe a hundred characters, so you're well under a thousandth of a dollar on the cheap tier. It's closed, though. Every top Hebrew voice in this bake-off is behind an API. The best open option is a generation behind.



Corn

And the voice characters? He went past the winners.



Herman

He did, and the notes are lovely. ElevenLabs version four, in a female voice, lost to the leader in the arena, but he said it has a bit of style, a kind of vibe, and if he wanted a voice to mix into a podcast, that's the one he'd choose. It costs eight cents per thousand characters. The male voices of Gemini Live and ElevenLabs version four he described as a deeper, boomy announcer voice, the kind popular in Hebrew TV adverts. Great for a phone tree, too flat for narration. ElevenLabs version four male, Brian, he called native-sounding with a single slightly non-native R in "baboker", in the morning. And Gemini three point eight Flash Lite male, given nikud, had an edge: at points, meaning is carried beyond the words.



Corn

Meaning carried beyond the words. A text to speech voice with subtext. I feel judged.



Herman

Don't be. Hebrew has a lot of meaning that sits above the words, which is why prosody is the hard part and the easy part to notice.



Hilbert

Olivetti Lettera thirty-two. Ninety-four pounds, secondhand, and the E sticks. It sits on the hall table and I still type the envelopes on it.



Corn

Ninety-four pounds for a sticky E. Mine has three.



Hilbert

When the county rates office addressed the bill to Flumington, I wrote to a Mr Dacre in the collections department. Said the name is Flumingtop, stress on the first syllable, and I would not pay until the bill said it right. Fourteen months. I put a pencil mark over the vowel on every cheque stub, so they could see which part gets the weight. Herman, you spent the better part of the episode telling us nikud leaves out the stress. It took me one letter to Mr Dacre.



Herman

A pencil mark over the vowel is more or less what Phonikud adds. The diacritizer gives you the vowels, a second step adds the stress, and that's exactly the layer that was missing from the old coach voice.



Corn

It took Mr Dacre fourteen months to fix one name. The models are working at speed.



Hilbert

He fixed it in the end. I'd asked him whether a machine could be told which part of a name gets the weight, and he said no. So, can it?



Herman

Yes, if somebody marks the stress somewhere the model can read it. If nobody does, it guesses. The best one guesses well, most of the time.



Hilbert

Then I was right to use pencil. Over the vowel, not in the margin. A margin is a smudge.



Corn

Same problem for Daniel, though. A pencil mark is easy to read, and nobody has written stress into nikud for a hundred sentences of Tatoeba.



Herman

Which is the thing the bake-off leaves open. Stress marking is first, because the points don't carry it and the models only partly infer it. Sef-ER and yel-ED came out of a system that had every vowel correct, so a better diacritizer on its own wouldn't have saved the old voice. The research groups working on this keep changing what they call the problem. Phonikud started as grapheme to phoneme conversion for Hebrew and ended up as a paper about underspecification, meaning the page just doesn't contain enough information to say the word. The same people have a newer approach, called ReNikud, which skips hand-labelled vowels and uses a speech recognizer to label thousands of hours of audio instead. Whether that code is out yet I haven't seen confirmed.



Corn

And the open models? You said the gap was a generation.



Herman

The open ones are fine-tunes on small corpora, mostly one speaker, mostly non-commercial licences. The open Hebrew line, Phonikud, ReNikud, Blue V three, is a handful of people, and Blue V three only came out four days ago. The distance between that and a closed model backed by a language model that's read the internet is a hundred points on the arena, as of the last time we looked. Closing it needs data nobody has released yet, or a cleverer way round the vowels. There's a Hebrew University paper from two years ago on diacritic-free Hebrew TTS over discrete speech units, an approach built on exactly that idea. I've only read the abstract.



Corn

Licences count too, surely. A voice you can't ship is a demo.



Herman

Yes. Phonikud's code is CC-BY but the training data is non-commercial. SASPEECH is non-commercial, and Kokoro Hebrew inherits that. So even where the weights are open, you may not be allowed to put the result in a product. The closed APIs are the ones with clean terms, which is a strange thing to be grateful to a closed API for. Corn, you'd be the one picking. Daniel hands you an app with a Listen button. Do you take the three cent voice at the top or the one cent voice a bit flatter?



Corn

The cheap one, and I'd listen to it on a bad phone speaker first. If he's standing in a kitchen reading nikud off a screen, he needs a clear voice more than an expressive one. The flatter voice could even help, as long as the stress lands in the right place.



Herman

That's roughly where Daniel landed on value. He just kept the better voice for the part where he's the listener. And there's another gap, in how the voices are judged. The crowd arena gives you an Elo number, a rank, and it can't tell you which error happened. The Phonikud group built audio to phoneme models for automatic evaluation for that reason. Daniel's per-clip reasons, osheh for oseh, hi for i, are the other half of it.



Corn

And the one that matters to Daniel personally?



Herman

Code switching. His sentences in the benchmark are in Hebrew, with Hebrew loanwords sitting in them. His real project is speech where somebody says half of a sentence in each language. I haven't seen a Hebrew English mixed benchmark for TTS anywhere. Deepdub, the Tel Aviv company, claims it handles it. Taking that at its word, that's a claim from a vendor that also claims to be number one on the arena.



Corn

Which is the same position Larry's spray is in.



Herman

Larry's spray comes with a certificate of authenticity. Deepdub's is a leaderboard page that contradicts itself by a couple of points.



Corn

Daniel's bake-off is a few dollars of compute and a pair of ears, and it did what the commercial pages didn't. Told us which voice reads "I would do it again" without asking a question at the end.



Herman

And it told him what to put in a button. Gemini three point one Flash, Kore, with checked nikud. The old voice is retired.



Corn

What stays with me is the pencil mark. Everything in this field is fighting over one little mark above a vowel, and the open models still can't draw it for themselves.



Herman

And whoever builds the first open voice that puts it in the right place, unprompted, on plain text, would close most of that hundred-point gap. That's the one to watch.



Corn

Thanks to our producer Hilbert Flumingtop for keeping all twelve of those clips in the right order. This has been My Weird Prompts. If you want the full bake-off, the leaderboard and Daniel's blog write-up, find us at my weird prompts dot com.



Herman

See you tomorrow.