



ASR WATCH · EDITION 2

The base model moment for speech

The top of the speech leaderboard barely moved this week. The action was underneath: smaller models, better handling of names, and turn-taking built in.

BY MAGAZINER · OCTOBER 2026

IN SHORT

In one week, three small labs shipped models built on NVIDIA's open Parakeet v3, each 156 to 253 MB, each claiming to match or beat the 2.5 GB original on its own tests. That is the pattern text models went through: a strong open base, then many small teams specialising it. Speech just got there. Meanwhile accuracy at the top barely moved, and the useful gains came from word lists for names, extras like speaker labels

SPEECH RECOGNITION JUST GOT its base model moment. In one week, three small labs shipped models built on NVIDIA's open Parakeet v3. Each is between 156 and 253 MB. Each claims to match or beat the 2.5 GB original on its own tests.

Text models went through this a while ago. A strong open model appears, and then many small teams take it and specialise it: for one language group, for one device, for one kind of noise. Speech is there now. And while that happened underneath, the top of the accuracy table barely moved. My argument this week is that the useful work in speech has moved below the leaderboard.



1

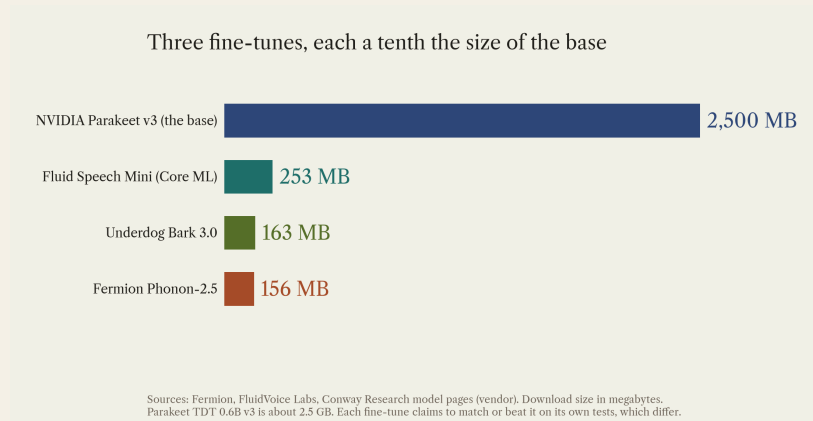
ONE BASE, THREE FINE-TUNES

Last week the on-device story was compression. Fermion had squeezed Parakeet v3 down to 164 MB with Phonon-2. This week, compressed Parakeet models are being trained past their base.

Fermion's Phonon-2.5 is 156 MB and keeps all 25 European languages. It adds automatic language detection and word timestamps, and runs on CPU, Apple silicon and CUDA. Fermion reports an average error of 11.71% on the FLEURS test set, against 13.26% for the full Parakeet v3 (vendor). Greek goes from 36.11 to 20.83 (vendor). That jump is big enough that I would want to see someone else reproduce it.

Fluid Speech Mini from FluidVoice is English only, a 253 MB package for Apple devices. On its own average over eight English test sets it scores 4.54, against 4.86 for Parakeet v3 (vendor). The more interesting number is on audio with other people talking in the background: 10.9 against 20.4 (vendor). FluidVoice also says the model does worse than larger ones on traffic noise, which is a nice bit of honesty.

Underdog Bark 3.0 from Conway Research is English and 163 MB. It ties Parakeet v3 on a seven-set average: 4.937 against 4.941 (vendor).



So which one is best? You can't tell. Bark averages seven English sets, Fluid averages eight, and Phonon uses FLEURS. All three ran their own tests. None of these numbers can be ranked against each other.

The more fine-tunes there are, the more you need to test on your own audio.

That is the catch with a base model moment. Choice gets better and comparison gets harder. If you want transcription on a laptop or a phone, or cheap bulk transcription on CPUs, the question is no longer “which model”. It’s “which fine-tune of Parakeet, and what audio was it tested on”.



2

NAMES MATTER MORE THAN ERROR RATE

Overall error at the top is now within a point or so across vendors. What still differs is whether the model gets the names right: the people, the products, the places. A transcript that is 96% right but spells the guest's name wrong every time is not

much use.

The cheapest fix is a word list. Mistral's new [Voxtral 3](#) takes one, and Mistral says recall of key phrases on the Earnings-22 test set goes from 81.4% without context to 94.6% with it (vendor).

A word list lifts key-phrases recall by 13 points



Source: Mistral, Voxtral 3 announcement (vendor). Earnings-22 key-phrase recall, percent, higher is better.

That 13-point gain is bigger than any change in overall error rate this week, and it costs you a list you probably already have. IBM went the same way. [Granite Speech 5.0](#) accepts up to 2,000 keywords, ships under Apache 2.0, and scores 5.06% on the official Open ASR results. If audio can't leave your own systems, that is the one I would test first.

The way I see it, the order of work has changed. Before you fine-tune anything, send the names.



3

THE EXTRAS ARE BECOMING THE PRODUCT

Voxtral 3 also shows where speech services are going. It bundles speaker labels, the word list and end-of-turn detection in one service, at \$0.0035 to \$0.004 a minute. On the independent [Artificial Analysis board](#), Voxtral Small scores 2.1% and Mini 2.2%, third and fifth.

But read the docs too. Mistral’s own [documentation](#) says there are no word timestamps and no handling of overlapping speech. If you build chapters, clips or aligned transcripts, that matters more than a tenth of a point of error.

Turn-taking is the other extra moving inside the transcriber. Voxtral 3 detects the end of a turn natively. ElevenLabs added turn prediction to its realtime service. Liquid AI’s LFM3-Audio and NVIDIA’s voice agent start preparing a reply while you are still speaking. A year ago that was a separate component bolted on after the words came out.



4

SPEECH ON A PHONE, FOR REAL

One developer packaged an EmbeddingGemma 2 based reader for Android phones ([model card](#)). On a Galaxy S26, a 54-minute audiobook took 71 seconds with the encoder on the NPU, about 46 times faster than real time. The card itself says one recording is “indicative only”.

The brittleness shows on LibriSpeech: 4.21% on the clean test set and 10.42% on the harder one. It’s English only, has no punctuation and was trained only on read audiobooks. Fast and cheap, then, but not yet ready for noisy real-world audio.



5

ONE ODD IDEA

[oruk-loop-153M](#) does something I hadn’t seen in speech. Instead of stacking many different layers, it runs one encoder block 24 times and one decoder block 28 times. It scores 7.12% on seven English test sets (vendor), though its own card says the

checkpoints were picked on those same sets. The card is also clear that looping saves storage, not compute: it isn't faster. Still, a model you store at 153M parameters and run deep is a neat trick for small devices.



6

WHAT TO DO WITH THIS

Accuracy at the top barely moved this week. Underneath, three things did. Open speech models now have a base to build on. Word lists became the cheapest accuracy you can buy. And speaker labels and turn-taking are moving into the transcriber itself.

So if I were picking a speech model today, I would start from a fine-tune of a strong open base, give it my list of names, check that it returns the timestamps I need, and test all of it on my own audio. The leaderboard can tell you who is close. Only your audio can tell you who is right. ■

About this series. ASR Watch is a weekly look at speech recognition. Edition 2 covers 21 to 28 October 2026. “(vendor)” means a company reported the figure about its own model; “(X post)” means one post on X that could not be checked further.

Sources: Fermion, [Phonon-2.5](#); FluidVoice Labs, [Fluid Speech Mini](#); Conway Research, [Underdog Bark 3.0](#); Mistral, [Voxtral 3 announcement](#) and [speech-to-text docs](#); IBM, [Granite Speech 5.0](#); [Artificial Analysis speech-to-text board](#); [EmbeddingGemma 2 Transcribe LiteRT model card](#); Oruk AI, [oruk-loop-153M](#); Liquid AI, [LFM3-Audio-2B](#).

Illustrations and charts were made for this article. Set in Libre Caslon and Josefin Sans.