Case Study: Replacing a Cloud Dictation Product in Three Days — and Measuring the Result

FirstPass Systems — Production-Grade AI for Manufacturing & Operations. Anthony Ford, founder. A founder build: the tool described here was built for daily use in the founder's own workflow, then benchmarked head-to-head against the commercial product it replaced, using that product's own usage logs. Every number below is measured, not estimated. The commercial product is anonymized.

The Challenge

The founder dictates constantly — notes, analysis, correspondence — through a well-funded cloud dictation product at roughly $180 per year. Two problems surfaced in its own logs. First, capacity: the subscription tier allowed 2,000 words per week, while actual measured usage ran 3,976 words per week — the plan ceiling was half the real need. Second, custody: every spoken word traveled to the vendor's cloud, and the product's local database was found to retain 152 audio recordings of the founder's voice, alongside database fields for capturing on-screen text from the applications being dictated into.

The Approach

FirstPass built a replacement in three days: hold a key (or a mouse button), speak any length, release — transcribed text lands at the cursor in any application. The design decision that matters: **everything runs on-device.** Speech recognition is a quantized open-source model on the local machine's own silicon; the language-model cleanup pass that strips filler words and false starts is a second on-device model, invoked only when the transcript actually needs it. Audio is discarded the moment it is transcribed. Nothing touches a network.

Reliability was engineered production-line style, with each failure mode found in live use met with a countermeasure:

- **Recognition-loop hallucination:** speech models famously repeat phrases endlessly over background noise (an air conditioner triggered a 100-repetition loop). Countermeasures: context conditioning disabled, a deterministic repetition collapser, and noise-floor-aware silence trimming. Verified against the recorded failure.
- **Silent capture loss:** a wireless microphone that goes to sleep can deliver near-silence. The system detects it, names the failed device, and fails over to a live microphone — a spoken word is never silently discarded.
- **Latency at length:** long dictations are transcribed in the background while the speaker is still talking, chunked at natural pauses. Wait time no longer grows with dictation length.

The Result — Measured Head-to-Head

The incumbent product logged its own end-to-end latency on all 574 of the founder's dictations over eleven weeks. The comparison:

Measure	Cloud incumbent (its own logs)	FirstPass build
Latency, short dictations (under 15 s)	0.61 s	~1.3 s
Latency, medium (15–45 s)	1.13 s	~1.3 s
Latency, long (over 45 s)	1.53 s and climbing with length	~1.3 s, flat at any length

Measured example	—	54-second dictation, 1.3 s wait
Word limit	2,000 per week	None
Recurring cost	~$180 per year	$0
Where speech is processed	Vendor cloud	On-device only; audio discarded after transcription
Hardware required	Vendor datacenter	A consumer laptop with 8 GB of memory

The honest reading, both directions: the cloud product remains faster on short bursts — its datacenter beats a laptop in a sprint. On long-form dictation, the three-day build matches or beats it, with no caps, no subscription, and no audio leaving the machine.

Why This Matters for Manufacturers

This build is the proof-of-engine for a plant-floor question: what happens when the person who just watched a line go down can simply *say* what happened — and it becomes a structured record? The same on-device speech stack, tuned to a plant's vocabulary of machine names, part numbers, and reason codes, is the basis of the FirstPass voice capture offering — with the same custody property: spoken words processed inside the building, never in someone else's cloud. It is also a working demonstration of the FirstPass premise itself: commercial software that limits your operation is not a fact of life. Sometimes it is three days of focused engineering.

Anthony Ford · FirstPass Systems · Thorntown, Indiana · anthony.ford@firstpasssystems.com