

Streaming ASR Provider Selection Worksheet

Whisper · Deepgram Nova-3 · AssemblyAI Universal-Streaming — pick by four questions

1 · Four questions, in order

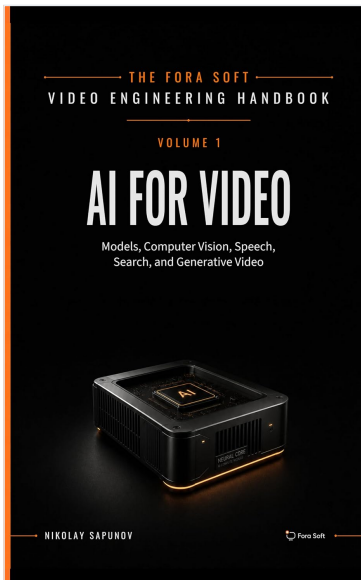
	Question	If yes
Q1	Must the audio stay on your own servers? (medical, legal, data-residency)	YES → self-hosted Whisper
Q2	Voice agent, or just a transcript? (does it need to know when to speak?)	AGENT → Deepgram / AssemblyAI
Q3	Do words need to appear in under ~300 ms? (instant captions / responsive UX)	YES → hosted API
Q4	Very high, very steady volume, plus an ML team to run a GPU fleet?	YES → self-host can win on cost

2 · The three options (vendor-reported, 2026)

Option	Latency	Price	Langs	Best when
Whisper (self-host)	~3.3 s avg	GPU only	99	On-prem; very high scale
Deepgram Nova-3	< 300 ms	~\$0.46/hr	45+	Lowest latency; broad features
AssemblyAI Universal-Streaming	~300 ms (immutable)	\$0.15/hr	99+	Voice agents; immutable transcripts

3 · Pre-flight checklist before you ship

- ☐ Decided exactly which results are **final** (immutable) vs **partial** (revisable) — and act only on finals for irreversible actions.
- ☐ Measured latency and word error rate on **your own audio**, not vendor demo clips.
- ☐ Confirmed language coverage matches your real user base (not just English).
- ☐ Checked data-residency / compliance: can the audio legally leave your servers?
- ☐ If a voice agent: end-of-turn detection chosen (native in Flux / Universal-Streaming, or built yourself).
- ☐ Modelled cost at **real, peaky** volume — not steady-state — including idle GPU capacity if self-hosting.
- ☐ Extracted clean audio from the WebRTC / call pipeline (correct sample rate, mono, no clipping).
- ☐ Load-tested concurrent streams at expected peak before launch.



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This worksheet narrows Whisper, Deepgram and AssemblyAI down with four questions and a pre-flight checklist. Volume 6 covers the speech chapter in depth: ASR and diarization, TTS, model artifacts and deployment topologies, and the ML minimum you need to judge a vendor benchmark.

WHAT YOU WILL FIND INSIDE

- Streaming vs batch ASR and how partial results behave
- Self-hosted vs API deployment topologies
- Speaker diarization alongside transcription
- Reading word error rate claims on your own audio

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Adding live transcription or a voice agent to your product?

Tell us your languages, latency target and compliance rules; we will benchmark the providers on your audio and ship the ASR pipeline. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)