

Real-Time Translation Architecture Decision Checklist

Cascade vs end-to-end, the 2-3 second window, the WebRTC sidecar, and the 2026 system comparison

1 · The four 2026 systems at a glance

System	What it is	How you get it	Languages	Best when
Google Meet	Built-in feature	Subscription	EN ↔ 5 langs	Everyday meetings; voice like yours
OpenAI translate	Rented model	API (~\$0.034/min)	70 in / 13 out	Build a product; dynamic per-speaker voice
SeamlessStreaming	Open-weights	Self-host	~100 input langs	Open baseline; CC-BY-NC (non-commercial)
Enterprise RSI	Contracted service	Contract	6,000+ combos	High-stakes events; human in the loop

Google Meet, OpenAI & Seamless are end-to-end engines; enterprise platforms (KUDO, Interprefy, Wordly) keep a human interpreter available for accuracy AI alone can't yet guarantee. Language counts & prices read May 2026; re-verify at use.

2 · Cascade or end-to-end?

Criterion	Cascade (3 models)	End-to-end (1 model)
How it works	Chain ASR → MT → TTS, readable text between each	One model: audio in language A, audio out language B
Latency	~1-2 s; adds up across 3 hand-offs	~400-700 ms; fewer hand-offs
Debuggability	High — inspect the transcript at each stage	Low — no readable middle to inspect
Voice	Voice flattened to text and back	Keeps tone, prosody, the speaker's voice
Lock-in	Swap any component, any vendor	One vendor's take on all three
Choose when	Audit trail, regulated, debuggability	Low latency & a natural voice matter most

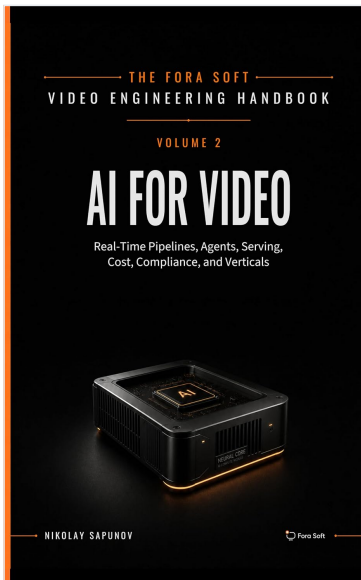
3 · Pre-flight before you ship a real-time translation feature

- ☐ Named the **format**: spoken translation (a voice) or translated captions (subtitles), or both, user's choice?
- ☐ Picked an **architecture**: cascade for control/debuggability, end-to-end for low latency + a natural voice.
- ☐ Set a **latency target** in the 2-3 second window — faster is hard to understand, slower breaks conversation.
- ☐ Decided the **speed-vs-accuracy** point: wait longer for accuracy, or commit faster for flow — you can't have both.
- ☐ Confirmed the delay stays **under ~2 seconds** in back-and-forth calls, or callers will start talking over each other.
- ☐ Chose one of the **four kinds of system**: Google Meet, OpenAI gpt-realtime-translate, SeamlessStreaming, or enterprise RSI.
- ☐ Verified **language coverage** for the pairs you must support, including structurally hard ones (e.g. verb-final German).
- ☐ Planned the **transport**: a sidecar taps the SFU's audio, translates, and feeds per-language tracks back to listeners.
- ☐ Budgeted **one session per target language**, not per participant (3 languages = 3 sessions, fanned out by the SFU).
- ☐ Avoided **word-for-word translation**: chose a model that translates meaning, so idioms and sarcasm don't become nonsense.
- ☐ Secured **consent** to clone a participant's voice — joining the call is not automatically consent (see NO FAKES Act).
- ☐ Added **provenance & safety**: watermark machine-generated audio and detect added toxicity at the output.
- ☐ Kept a **human fallback** for high-stakes content (board meetings, earnings calls, clinical consultations).

Companion to: 'Real-Time Multilingual Speech Translation In Calls'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/real-time-multilingual-speech-translation-in-calls

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 7 OF 7

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

From Real-Time Pipelines and Agents to Serving, Cost, Compliance, and Verticals

By Nikolay Sapunov, CEO at Fora Soft · 1,221 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0GX3DGBH4

This checklist compares four 2026 translation systems, cascade vs end-to-end, and the pre-flight checks before launch. Volume 7 covers this in depth: live captions and translation under a real-time budget, where AI fits into a WebRTC pipeline, serving and cost per stream-hour, and voice and biometric regulation.

WHAT YOU WILL FIND INSIDE

- Live captions and translation in calls
- Serving cost per stream-hour
- Where AI fits into a WebRTC pipeline
- Voice cloning, consent and regulation

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

Choosing between cascade and end-to-end translation for your calls?

We will prototype both on your language pairs, measure delay and accuracy, and ship the architecture that fits your users. 250+ video products shipped since 2005.

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