

Real-Time Speech Translation — Earbuds & WebRTC — Cheat Sheet

AI for Video Engineering

1 • The four-step pipeline

Hear → ASR (speech to text) → MT (translate) → TTS (text to speech) → deliver.
Same pipeline whether the audio is from earbuds or a video call.

2 • 'Translation earbuds' = an audio endpoint in front of the pipeline

On-device (Apple Live Translation, iOS 26): private, offline, hardware-capped.
Phone + cloud (Pixel / Galaxy): Gemini-class, 70+ languages, needs a network.

3 • Cascade vs direct (how to build the brain)

Cascade (ASR→MT→TTS): readable text → captions, glossary, debug, audit. More delay.
Direct S2ST (Seamless / GPT Realtime / Gemini Live): lower delay, keeps the voice.
Regulated + caption-heavy → cascade. Latency-critical + natural voice → direct. Often both.

4 • Why it rewrites itself + the latency budget

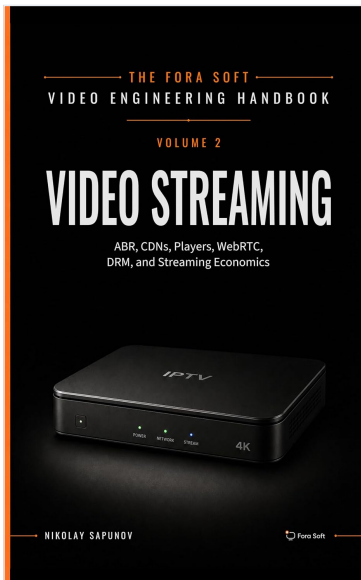
Languages reorder words → early translation is a guess. REPLACE the line until final.
Stay ~2–3 s behind (ear-voice span / wait-k). Speech path waits longer than captions.
Budget: ~1 s machine (TTS biggest) + ~2–3 s wait. <800 ms feels live; ALC ceiling 3–5 s.

5 • In a WebRTC call: translate once at the SFU, fan out per language

Tap speaker audio at the SFU, VAD-gate it, translate once per listener language.
Captions → data channel (RFC 8831). Audio → injected Opus track (RFC 6716), duck original.
Cost ~ languages × minutes: 3 langs × 60 min × \$0.12 = \$21.60/hr. VAD gates the rest.

6 • Build-vs-buy checklist

- ☐ Group call you record / audit? → cascade at the SFU (text for captions + record).
- ☐ Two-person, natural-voice feel? → direct S2ST model, expressive variant.
- ☐ Self-host the call? → mediasoup / Janus / LiveKit (own the tap + track injection).
- ☐ Pick ASR + MT + TTS (cascade) or one S2ST model; swap on first-chunk latency + accuracy.
- ☐ <60 concurrent streams → cloud API wins; far above → self-host (compliance decides).
- ☐ Always: VAD-gate, add a glossary for proper nouns, plan voice-clone consent + a record.



GO DEEPER · VOLUME 3 OF 7

Video Streaming: A Complete Guide to Video Delivery

From ABR and CDNs to Players, WebRTC, DRM, and the Economics of Streaming

By Nikolay Sapunov, CEO at Fora Soft · 1,182 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH76ZYWC

This sheet covers the speech translation pipeline, earbuds vs calls, cascade vs direct and SFU fan-out per language. Volume 3 covers the WebRTC side in depth: SFU topologies and track injection, players and recovery, QoE monitoring, and the economics of delivering every extra language track.

WHAT YOU WILL FIND INSIDE

- WebRTC SFU topologies and track injection
- Data channels for live captions
- QoE monitoring of real-time calls
- Delivery cost per extra stream

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

Want live speech translation inside your video calls?

We will add per-language translation at your SFU, deliver captions and dubbed tracks, and keep the cost per minute predictable. 250+ video products shipped since 2005.

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