Speech-to-Speech Architecture Decision Checklist

Cascade vs end-to-end, the 800 ms budget, transport choice, turn-taking, and the 2026 system comparison

1 · The three 2026 systems at a glance

System	Job	Host / rent	Connection	Cost shape	Best when
OpenAI Realtime API	Conversation	Rent (API)	WebRTC/WS/SIP	~$0.18-0.46/min	Production voice agent, phone
Gemini Live	Conversation	Rent (API)	WebSocket	~10x cheaper input	High-volume on a budget
SeamlessM4T v2	Translation	Host (open)	DIY	GPU + power	Open translation baseline

OpenAI & Gemini are end-to-end conversation engines; SeamlessM4T is an open-weights translation model under a CC-BY-NC (non-commercial) licence — clear licensing before shipping. Per-minute & per-token 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 -> LLM -> TTS, readable text between each	One model: audio in, audio out, no readable middle
Latency	Adds up across 3 handoffs; grows with model size	Lower — no handoffs
Debuggability	High — inspect the transcript at each stage	Low — no readable middle to inspect
Naturalness	Voice flattened to text and back	Keeps laughs, tone, interruption, accent
Lock-in	Swap any component, any vendor	One vendor's take on all three
Choose when	You need control & debuggability	The conversation must feel human & fast

3 · Pre-flight before you ship a speech-to-speech feature

- [ ] Named the job: conversation (answer in same language) or translation (say it in another language)?
- [ ] Picked an architecture: cascade for control/debuggability, end-to-end for low latency + human feel.
- [ ] Set a voice-to-voice budget: target < 800 ms; remember the model is only ~half (≈ 500 ms TTFB in US).
- [ ] Measured end-to-end voice-to-voice, not just the model's response time, from your users' real locations.
- [ ] Chose the transport: WebRTC for browser/mobile device audio, WebSocket for server pipelines, SIP for phone calls.
- [ ] Tuned end-of-turn detection (VAD) on real, messy user audio — accents, noise, mid-sentence pauses — not a quiet demo.
- [ ] Built interruption handling (barge-in) as a first-class feature: stop output, flush the queue, restart listening.
- [ ] Costed your minutes/month against both OpenAI and Gemini price sheets, including prompt-caching discounts.
- [ ] Switched on prompt caching where available (OpenAI caches audio input ~80x cheaper than fresh input).
- [ ] Confirmed language coverage for the languages you must support, including mid-sentence switching if needed.
- [ ] Checked the licence if self-hosting: SeamlessM4T is CC-BY-NC — commercial use needs a separate arrangement.
- [ ] Verified data residency and privacy: where does the audio travel, and does that meet your regulatory posture?