Gemini & Google voice — ultra research juli 2026

> Samengesteld door Kimi, voortbouwend op Conductor's eerdere research (4 verslagen, juni 2026).
> Aangevuld met verse web-research van 25 juli 2026. Alle bedragen en claims hebben een bron.
> Onzekere punten zijn eerlijk gemarkeerd met "let op".

In het kort (de 5 belangrijkste dingen)

1. **Google heeft op 21 juli 2026 nieuwe efficiëntie-modellen uitgebracht** (oa Gemini 3.6 Flash). Goedkoper, sneller, gebouwd voor agents. Geen revolutie, wel direct bruikbaar.
2. **Voor onze podcasts verandert er bijna niks.** Sulafat blijft de stem, dat is beslist. Er is wel een nieuwer TTS-model (3.1) dat we eenmalig kunnen testen.
3. **De Gemini Live API (echt praten met een AI) is nu serieus goedkoop**: ongeveer 2,3 cent per gespreksminuut. Dat is 3 tot 5 keer goedkoper dan ElevenLabs.
4. **Toch niet klakkeloos overstappen met Roos** (de telefoon-stem). ElevenLabs is nog steeds beter in Nederlandse stemkwaliteit en telefoon-koppeling. Eerlijk advies: hybride test.
5. **Mijn voorstel**: 3 kleine proeven draaien (hieronder uitgewerkt). Geen grote migratie, wel meten.

---

1. Wat er al bekend was (Conductor's research, juni)

Conductor heeft dit onderwerp op meerdere pagina's uitgezocht. Die research heb ik gebruikt, niet opnieuw gedaan:

- Gemini Omni — analyse (14 juni)

- Gemini API Referentie (12 juni)
- Gemini Live in de auto (18 juni)
- Nieuwe Gemini AI-modellen — opdracht via Jarvis (23 juni)

Wat daar in juni al uitkwam:

- **Gemini Live** (realtime praten via WebSocket) ondersteunt Nederlands, kost ongeveer $1 per uur. Dat was toen al 7 tot 12 keer goedkoper dan OpenAI of ElevenLabs.
- De TTS-modellen kenden al 30 stemmen, waaronder onze Sulafat.
- Voor beeld en video was de conclusie: Omni handmatig, Veo 3.1 via de API.
- Open punt uit juni: wat doen de nieuwe modellen met kosten, snelheid en kwaliteit? Dit verslag beantwoordt dat.

---

2. Wat er nieuw is sinds juni (juli 2026)

De modellen van 21 juli

- **Gemini 3.6 Flash** (GA): $1,50 / $7,50 per miljoen tokens (in/uit). Verbruikt circa 17 procent minder output-tokens dan 3.5 Flash. Sterker op coding en agent-taken.
- **Gemini 3.5 Flash-Lite** (GA): $0,30 / $2,50 per miljoen. Snelste van zijn generatie (350 tokens per seconde). Ideaal voor bulk-werk: samenvatten, sorteren, routeren, subagents.
- Let op bij migratie: de oude instellingen temperature, top_p en top_k worden nu genegeerd.

De Live API (echt praten, stem naar stem)

- Werkt via WebSocket. Audio erin op 16 kHz, eruit op 24 kHz. Meer dan 70 talen, dus ook Nederlands.
- Nieuwe trucs: **affective dialog** (de stem reageert op je emotie) en **proactive audio** (het model kiest zelf of het antwoordt).
- Nieuwste model: gemini-3.1-flash-live. Kosten: ongeveer 2,3 cent per gespreksminuut. Reactiesnelheid in de praktijk: 250 tot 500 milliseconden.
- Leuk feitje: Google Voice Search draait hier nu op, in meer dan 200 landen.

De TTS (tekst naar spraak, onze podcast-stem)

- Nieuwste model: **gemini-3.1-flash-tts-preview**. Enige met streaming, en batch is 50 procent korting.
- Onze huidige gemini-2.5-flash-preview-tts blijft gewoon bestaan: circa 1,5 cent per minuut audio.
- Bekende beperking (wij merkten dit zelf ook): de kwaliteit zakt na een paar minuten. Daarom knippen wij de tekst in stukken van 1200 tekens. Dat blijft nodig.

---

3. De eerlijke vergelijking: wie maakt de beste voice-agent?

OPTIE	TARIEF PER GESPREKSMINUUT	SNELHEID	NL-STEMKWALITEIT
Gemini Live	circa $0,023	250-500 ms	goed, maar geen vaste "eigen" stem
ElevenLabs (Roos)	$0,08 tot $0,12	500-800 ms	de beste, inclusief voice cloning
OpenAI Realtime	$0,06 tot $0,15	circa 320 ms	geen voordeel voor NL

Wat dit betekent, eerlijk gezegd:

- **ElevenLabs wint nog op kwaliteit.** De Nederlandse stemmen klinken het natuurlijkst, en Roos is inmiddels een herkenbare stem voor de salon.
- **Gemini wint op tarief en snelheid.** Bij veel gesprekken (duizenden minuten per maand) scheelt dat 3 tot 5 keer.
- **OpenAI is voor ons nu de minst aantrekkelijke.** Duurder dan Gemini, geen Nederlands voordeel.
- Let op: er is nog geen onafhankelijke test van de nieuwste Gemini-modellen speciaal voor Nederlands. Daarom zeg ik: zelf proeven, niet geloven.

4. Wat betekent dit per Haarvisie-product?

- **Blog-podcasts (Sulafat)**: niks veranderen. Goedgekeurd recept blijft staan. Wel eenmalig het nieuwe 3.1-TTS-model testen op één podcast: beter of niet, dan weten we het.
- **Roos aan de telefoon (ElevenLabs)**: niet overstappen. Wel een kleine hybride proef overwegen: simpele gesprekken (afspraak bevestigen, openingstijden) via Gemini Live, de echte merkgesprekken bij Roos houden.
- **Tekst-agents en bots (blogs, research, queue-werk)**: hier is de grootste winst. De nieuwe Flash-modellen zijn goedkoper én sneller. Test 3.6 Flash op één bot-taak en meet het verschil.
- **Jarvis en spraak in de auto**: de juni-conclusie blijft kloppen (Siri-shortcut nu, bel-bridge later). De nieuwe Live API maakt die bridge alleen maar aantrekkelijker.

---

5. Mijn voorstel: 3 kleine proeven

1. **Podcast-proef**: dezelfde podcast 1 keer met het nieuwe TTS-model maken. Jij luistert en kiest. Kost circa 2 cent.
2. **Agent-proef**: één terugkerende bot-taak op Gemini 3.6 Flash draaien en kosten plus snelheid meten tegenover nu.
3. **Voice-proef (optioneel)**: een minimale Gemini-Live-testbot die alleen openingstijden en afspraak-bevestiging doet. Puur om te horen hoe het klinkt in het Nederlands.

Geen grote migratie. Eerst meten, dan beslissen. Jij keurt elke proef goed voordat er iets live gaat.

Bronnen

- Gemini 3.6 Flash — Coursiv
- AI-nieuws 22 juli 2026 — BuildFastWithAI

- Nieuwe Gemini-modellen — Dutch IT Leaders
- Live API — officiële docs
- Prijzen — officiële docs
- TTS — officiële docs
- Gemini 3.1 Flash Live — Implicator
- Gemini Live vs ElevenLabs — MindStudio
- Voice-agent tarieven per minuut — Ringlyn
- ElevenLabs ConvAI vs TRTC
- OpenAI Realtime tarieven — HackerNoon
- Gemini vs ElevenLabs stemtest — Podonos
- Speech-to-speech architectuur — Leadlock