

Stem-Conductor — back-end handleiding

Dit is de complete technische uitleg van de **live-stem-Conductor**.

Daarmee praat Jamal **live met Claude (de Conductor)** via zijn stem.

Net als de back-end-handleiding die we voor Jarvis hebben.

1. Wat is het + adres + login

De stem-Conductor is een webpagina.

Je opent de pagina, drukt op de microfoon, en praat.

Je ziet **live je eigen tekst** verschijnen.

Je ziet ook **het antwoord van Conductor** als tekst.

En Conductor **praat terug** met de Sulafat-stem.

Het brein is **Claude/Conductor mét tools** — dus mét geheugen én volle macht over Jamals computers.

Het is dus geen kale chatbot, maar de echte Conductor met handen.

Adres:

- **Op de MacBook:** `http://localhost:8765`
- **Op de iPhone** (via Tailscale): `https://conductor-macbook.tail2d7247.ts.net`

Login (Basic-auth):

- gebruiker: **jamal**
- wachtwoord: **conductor2026**

Belangrijk over het adres:

- Op de MacBook moet je **localhost** gebruiken, niet de https-link.
- Reden: MagicDNS staat op de MacBook UIT (ooit gezet voor de Telegram-IPv6-fix).
- Daardoor is de `.ts.net` -naam op de MacBook een dood adres (NXDOMAIN).
- Op de iPhone staat MagicDNS wel aan, dus daar werkt de https-link.

2. Architectuur (hoe het in elkaar zit)

De stem-Conductor heeft vier onderdelen.

A. De server (het hart)

- Bestand: `~/ .claude/scripts/praat-conductor-server.py`
- Gebouwd met de standaard Python `http.server` (geen extra framework).
- Draait op **poort 8765**.
- Bindt **alleen** op `127.0.0.1` (localhost) en op de **Tailscale-IP** `100.88.240.64` .
- **NOOIT** op `0.0.0.0` — dat zou een LAN-datalek zijn.
- Bovenop Tailscale zit ook nog **Basic-auth** (jamal / conductor2026).
- Dit is kritiek: via de server draait Conductor met volle uitvoer-rechten (shell).

B. De LaunchAgent (houdt het aan)

- Bestand: `~/Library/LaunchAgents/nl.haarvisie.praat-conductor.plist`
- Label: `nl.haarvisie.praat-conductor`
- Start de server met `/usr/bin/python3 ../praat-conductor-server.py` .
- `RunAtLoad = true` → start meteen bij inloggen.
- `KeepAlive = true` → start automatisch opnieuw als hij stopt.
- Logs: `~/ .claude/praat/server.log` (uitvoer) + `~/ .claude/praat/server.err` (fouten).

C. De motor (Anthropic Messages-API)

- Vroeger startte elke beurt de hele `claude -CLI` opnieuw op (9 seconden vaste boot).
- Nu praat de server **rechtstreeks met de Anthropic Messages-API** (via `urllib`).
- Functie `_anthropic()` doet de API-call.
- Model: **claude-sonnet-4-6** (snel genoeg voor stem).
- Geheugen van het gesprek zit in een globale lijst `MESSAGES` .

D. De hand (run_bash-tool)

- Conductor heeft één krachtige tool: **run_bash** .
- Daarmee draait hij elk shell-commando op Jamals MacBook (`bash -lc`).

- Functie `_bash()` voert het uit (timeout 120s, output afgekapt op 4000 tekens).
- Hiermee kan hij in de praktijk ALLES wat de gewone Conductor kan: ssh naar de fleet, scripts draaien, geheugen lezen, research via curl, websites openen.

Hoe een beurt loopt:

1. Pagina stuurt je tekst (en evt. cameraplaatje) naar `/ask_stream`.
2. `run_conductor()` praat met de Messages-API.
3. Wil Claude een commando draaien, dan loopt de **tool-lus** (max 24 rondes).
4. Antwoord wordt in zinnen geknipt en **per zin** als Sulafat-audio teruggestuurd.
5. Pagina speelt de audio gapless af in een wachtrij.

3. Waar staat wat (alle paden)

ONDERDEEL	PAD
Server-script	<code>~/.claude/scripts/praat-conductor-server.py</code>
Webpagina	<code>~/.claude/praat/index.html</code>
Secrets (API-keys)	<code>~/.claude/praat/.env</code> (rechten 600)
LaunchAgent	<code>~/Library/LaunchAgents/nl.haarvisie.praat-conductor.plist</code>
Uitvoer-log	<code>~/.claude/praat/server.log</code>
Fout-log	<code>~/.claude/praat/server.err</code>
Werkmap / cameraplaatje	<code>~/.claude/praat-workspace/</code> (camera = <code>cam.jpg</code>)
TTS-script (Sulafat)	<code>~/.claude/scripts/tts-gemini.py</code>
Telegram-verzender	<code>~/.claude/scripts/telegram-send.sh</code>
ffmpeg	<code>~/.local/bin/ffmpeg</code>
Geheugen-index	<code>~/.claude/projects/-Users-macbookjamal/memory/MEMORY.md</code>

ONDERDEEL	PAD
Voorkeuren-register	~/ .claude/projects/-Users- macbookjamal/memory/voorkeuren_jamal.md

In `.env` staan:

- GEMINI_API_KEY (voor de Sulafat-stem)
- ANTHROPIC_API_KEY (voor de motor)

Reden voor een 600-bestand: onder de LaunchAgent kan de Keychain locked zijn.

4. Veilig herstarten

Als de server hapert of je hebt het script aangepast: herstart **alleen de LaunchAgent**.

```
launchctl kickstart -k gui/$(id -u)/nl.haarvisie.praat-conductor
```

Verifiëren dat hij leeft:

```
curl 127.0.0.1:8765
```

Krijg je **401** terug, dan leeft de server (dat is de login-vraag).

NOOIT doen:

- NOOIT Claude/Conductor zelf herstarten of afsluiten.
- Alleen deze ene service kickstarten.

Logs bekijken bij problemen:

```
tail -50 ~/.claude/praat/server.log
```

```
tail -50 ~/.claude/praat/server.err
```

5. Functies (wat hij allemaal kan)

Geheugen + context-injectie

- `_load_context()` laadt bij het opstarten zijn geheugen in de system-prompt.

- Dat zijn: `~/CLAUDE.md` , `~/claude/CLAUDE.md` , `MEMORY.md` , `voorkeuren_jamal.md` , `reference_hermes_agent.md` , `reference_praat_conductor.md` en `reference_machine_namen.md` .
- Elk bestand max ~9000 tekens (samen ~33KB).
- Daardoor kent hij Jamal, Haarvisie, de regels, de projecten, de machines.
- Voor meer detail leest hij zelf dieper met `run_bash ( cat / grep` in de geheugenmap).

Volle macht via de shell

- Via `run_bash` : `research (curl)`, websites openen (`open`), ssh naar de fleet (imac-nl, coco-veranda, paco-spanje, de VPS), scripts draaien, secrets ophalen.
- Hij doet gewone dingen proactief, zonder te vragen.
- Alleen bij écht destructief/duur (verwijderen, geld, haarvisie.nl) vraagt hij eerst.

Sulafat-stem met streaming

- `tts_sulafat()` maakt de audio via `tts-gemini.py --voice Sulafat` .
- Output is ogg → ffmpeg → mp3 → base64 data-URI in de JSON.
- **Per-zin streaming** (`/ask_stream`): het antwoord wordt in zinnen geknipt (`split_sentences()`) en zin voor zin als audio gestuurd, zodat hij begint te praten zodra de eerste zin klaar is.
- Lukt de Gemini-stem niet, dan valt de pagina terug op de robot-browserstem.

Camera (vision)

- Knop  zet de camera aan.
- Bij een vraag wordt een snapshot meegestuurd als image in de user-message.
- De server schrijft het ook weg als `cam.jpg` , zodat Conductor ernaar kan kijken.

Transcript zichtbaar

- Je ziet live je eigen woorden én het antwoord van Conductor op de pagina.
- Geen voice notes: Jamal wil kunnen meelezen.

/end → analyse-verslag naar Telegram

- Bij de STOP-knop () , of als de pagina sluit, of na 30 minuten stilte: `publish_verslag()` .
- Die maakt een **analyse** van het gesprek (wat ging goed / wat kan beter / actiepunten).
- Publiceert het op verslag.haartips.nl en stuurt de **klikbare link naar Telegram**.

Zelflerend (voorkeuren)

- Zeg je "ik vind X goed/fout", dan schrijft Conductor dat zelf met `run_bash` weg.
 - Bestand: `voorkeuren_jamal.md` , met een datumregel onder de juiste kop.
 - Naast photo-feedback (Supabase) en Mnemosyne.
-

6. Changelog v1 → v11

- **v1** — Eerste werkende versie. Brein = `claude -p ... --continue` . Live-spraak via Web Speech API. Werkt op Chrome (Mac/Android). Sulafat-audio bewezen.
- **v2** — Jarvis-functies: live-indicator (groene stip), animatie-orb (luisteren=rood / denken=draait / praten=goud), pauzeknop, camera + vision. Mic-gotcha ontdekt (alleen op `https/localhost`).
- **v3** — Auto-verslag bij `/end` of 3 min stilte: samenvatting → `verslag.haartips.nl` → Telegram-link.
- **v4** — Sulafat-fix (`ffmpeg-PATH` onder `launchd`) + `GEMINI_API_KEY` uit `.env` . Latency aangepakt met `--strict-mcp-config` (geen MCP-boot).
- **v5** — Grote motor-wissel: `claude -p` (9s boot) vervangen door directe **Anthropic Messages-API** + `bash-tool`. ~10s → ~6s per beurt. `ANTHROPIC_API_KEY` in `.env` .
- **v6** — Per-zin streaming (`/ask_stream` , `ndjson`): TTS per zin, gapless audio-wachtrij in de pagina.
- **v7** — Geheugen-injectie (`_load_context`), zelfherstellende historie, verslag-fix met `Popen` (build loslaten), `do_HEAD` voor de live-indicator, volle macht expliciet in de system-prompt.
- **v8** — Klikbare HTTPS-link via Tailscale (`tailscale serve`) na "HTTPS Certificates" aanzetten.
- **v9** — iPhone-spraak: `record` → `/transcribe` (faster-whisper "small", NL), want iOS-Safari heeft geen Web Speech API. DNS-gotcha gevonden (MacBook = localhost).
- **v10** — Zelflerend werkt (`voorkeuren_jamal.md`), iOS-fix (forceer opname-modus op iPhone), Jarvis-lessen toegepast, playbook gepubliceerd.
- **v11** — STOP-knop + geheugen-fix + analyse-verslag.

De 'te ingewikkeld'-fix (lus 8 → 24)

- Symptoom: bij elke echte klus (bv. Homebrew + Node installeren) zei de stem steevast "Het werd te ingewikkeld, probeer het korter".
- Oorzaak: de tool-lus in `run_conductor._turn()` stond op `range(8)`. Een installatie/fleet-taak doet méér dan 8 bash-calls → de lus liep vol → kale foutmelding.
- Fix: de lus van **8 naar 24** gezet. Loopt hij nu toch vol, dan geeft hij het **laatste echte antwoord** terug in plaats van een fout.

De geheugen-wis-bug (v11)

- De idle-watcher flushte het geheugen al na 3 minuten stilte.
- Bij een pauze raakte Conductor dan het gesprek kwijt ("verse sessie").
- Fix: `IDLE_FLUSH` van 180 naar **1800 seconden (30 min)**.

7. Bekende problemen + fixes

#	PROBLEEM	OORZAAK	FIX
1	Microfoon "not allowed"	Mic mag alleen op een veilig adres (https of localhost); Tailscale-IP via http telt niet	MacBook = <code>http://localhost:8765</code> ; iPhone = de https-link
2	Sulafat valt weg (robotstem)	Service vond <code>ffmpeg</code> niet (kale PATH onder <code>launchd</code>)	<code>PATH ~/.local/bin:...</code> meegeven aan de TTS-subprocess + <code>GEMINI_API_KEY</code> uit <code>.env</code>
3	Loodzwaar traag (~10s/beurt)	Elke beurt startte de hele <code>claude -CLI</code> op (9s vaste boot, los van het model)	Motor vervangen door de directe Anthropic Messages-API
4	"Verbinding mislukt" blijft hangen (HTTP 400)	Na een hapering raakte de <code>MESSAGES</code> -historie corrupt	Zelfherstel: mislukte beurt eruit (<code>del MESSAGES[start:]</code>); bij 400 <code>MESSAGES.clear()</code> + 1x schoon opnieuw

#	PROBLEEM	OORZAAK	FIX
5	Geen geheugen ("kale" Claude)	Niets ingeladen	<code>_load_context()</code> : CLAUDE.md + regels + MEMORY + Hermes ingeladen; leest zelf dieper via <code>run_bash</code>
6	Verslag komt niet binnen	<code>build.py</code> op de VPS te traag (218 verslagen) → ssh-build timed out	Build loslaten met <code>subprocess.Popen</code> → Telegram-link gaat meteen, html volgt later
7	"offline" bovenin de pagina	De HEAD-ping werd niet afgehandeld	<code>do_HEAD</code> toegevoegd (geeft nu 200)
8	Gemini-stem hikt (geen audio, finishReason OTHER)	Transient Gemini-beperking	In <code>tts_sulafat()</code> tot 3x opnieuw proberen; faalt zacht → browserstem
9	Link werkt niet op de MacBook (NXDOMAIN)	MagicDNS staat op de MacBook UIT (Telegram-IPv6-fix)	MacBook = <code>localhost</code> , iPhone = de <code>.ts.net</code> -link
10	iPhone kan niet praten	iOS-Safari heeft geen (betrouwbare) Web Speech API	Forceer opname-modus op iOS: <code>record</code> → <code>/transcribe</code> → <code>faster-whisper</code>

8. Storing-zoekgids (stap voor stap)

Pagina laadt niet / "kan geen verbinding maken"

1. Leeft de server? `curl 127.0.0.1:8765` → moet **401** geven.
2. Niet? Herstart: `launchctl kickstart -k gui/$(id -u)/nl.haarvisie.praat-conductor` .
3. Check de fout-log: `tail -50 ~/.claude/praat/server.err` .

"offline" staat bovenin

- De HEAD-ping bereikt de server niet. Herstart de LaunchAgent (zie boven).

Microfoon werkt niet ("not allowed")

- Op de MacBook? Gebruik `http://localhost:8765` , niet de https-link.
- Op de iPhone? Gebruik de https-link (`conductor-macbook.tail2d7247.ts.net`).
- Op de iPhone gebruik je sowieso de **opnemen-knop** (tik = start, tik = stop).

Conductor praat met robotstem (geen Sulafat)

1. Check de log op `tts-fout : grep tts-fout ~/.claude/praat/server.log` .
2. Staat de GEMINI_API_KEY in `~/.claude/praat/.env` ?
3. Is ffmpeg er? `ls ~/.local/bin/ffmpeg` .
4. Vaak is het tijdelijk (Gemini hikt) — even opnieuw proberen.

Alles is traag

- ~5-6 seconden per beurt is ongeveer de bodem met Sulafat (de Gemini-stem is traag + wisselvallig).
- De motor (Claude-API) is snel; de stem is de bottleneck.
- Echt Jarvis-snel (~0,2s stem) zou ElevenLabs vereisen (andere stem + kosten) — Jamal koos bewust voor Sulafat.

"Het werd te ingewikkeld" bij een echte klus

- Dit hoort opgelost te zijn (tool-lus 8 → 24).
- Komt het terug bij hele grote taken: lus verder ophogen in `run_conductor._turn()` (`range(24)`), dan herstarten.

"Verbinding mislukt" blijft komen

- De historie was corrupt; de self-heal pakt dit normaal op.
- Lukt het niet: herstart de LaunchAgent (wist `MESSAGES` schoon).

Geen verslag binnengekomen na een gesprek

- De build op de VPS duurt ~1-2 min — niet te vroeg checken.
- De Telegram-link hoort wel meteen te komen.
- Check de log op `verslag-fout` .
- Test de link na een paar minuten: `https://verslag.haartips.nl/r-conductor-gesprek-...html` .

iPhone: link doet het niet

- Staat Tailscale aan + Connected op de iPhone?
- Staat "HTTPS Certificates" aan in de Tailscale-admin?
- Draait tailscale serve nog? Opnieuw: `tailscale serve --bg http://127.0.0.1:8765 .`

Algemene check-volgorde bij elke storing

1. `curl 127.0.0.1:8765 → 401 = leeft.`
2. `tail -50 ~/.claude/praat/server.log` en `server.err` .
3. `launchctl kickstart -k gui/$(id -u)/nl.haarvisie.praat-conductor` .
4. NOOIT Claude/Conductor zelf herstarten.