Build a voice agent for $0 on Google Cloud free tiers

> *"I want a voice agent answering callers or website visitors, but every platform wants a monthly fee before I've even proven the concept works for my business."*

Category: customer-communication | **Setup Time:** ~180 minutes | **Estimated Hours Saved:** 4 hrs/week | **Direct Software Cost:** $0 / month

SOLUTION SUMMARY

A browser-based voice agent on Google Cloud's free tiers genuinely runs at $0/month inside the quotas (GCP free-tier quotas checked 2026-10-21): Cloud Run's free request allowance hosts it, Speech-to-Text's free 60 minutes/month hears it, Text-to-Speech's free character allowance voices it, and a free-tier LLM answers. Three honest boundaries: GCP still requires billing enabled (set a $0.01 budget alert first), $0 covers a website demo — a real phone number and PSTN minutes are always paid — and past the free quotas it's a metered bill, not a cliff. Prove the concept at $0, then pay only for the channel that earns it.

REQUIRED TOOLS & DIRECT WHOLESALE UTILITIES

- **Google Cloud free tiers (Cloud Run + Speech-to-Text + Text-to-Speech + Gemini): (Free Tier / Wholesale)**
- **Vapi + Twilio (the paid production path): (paid)**

STEP-BY-STEP APPLICATION INSTRUCTIONS

1. **Open GCP, enable billing, set a $0.01 budget alert first**
   Free-tier services still require an active billing account. Before deploying anything, create a budget alert at the lowest threshold — if a quota slips, you hear about it in minutes, not on an invoice. New GCP customers may also qualify for a credit grant; eligibility varies and isn't required for the free tiers.

2. **Deploy the agent container to Cloud Run**
   The agent is a small websocket/WebRTC service: audio in → STT → LLM → TTS → audio out. Cloud Run's free request allowance covers demo traffic; scale-to-zero means it costs nothing while idle.

3. **Wire Speech-to-Text and Text-to-Speech**
   STT V1 carries a 60-minute monthly free allowance; standard TTS voices carry a large free character allowance. Stay on standard voices — the premium/neural voices are the paid meters.

4. **Point the brain at a free-tier LLM**
   Gemini's free tier handles the conversational turn. Keep the system prompt short and the tool list tiny — every extra token per call eats the same free quota faster.

5 Put it on your site as a browser call button — not a phone number

The $0 version lives behind a 'Talk to the agent' button on your site: visitors speak from their browser, no PSTN involved. The moment you want a real phone number answering, you've graduated to the paid path — and by then the demo has told you whether it's worth it.

6 Watch the three quota meters weekly

Cloud Run requests, STT minutes, TTS characters — each has its own free ceiling. A quiet demo site stays inside all of them; a promoted launch won't. That's the honest trade: $0 buys you proof, not production.

5-POINT QUALITY VERIFICATION PROTOCOL

[ ] **Carrier & Webhook Test:** Confirm webhook listener receives test payload and returns HTTP 200 within 500ms.
[ ] **End-to-End Simulation:** Execute live test event from personal device and verify SMS/email delivery.
[ ] **Anti-Spam & Rate Limit Check:** Verify repeated submissions do not trigger duplicate messages to clients.
[ ] **Direct Ownership Audit:** Confirm all API keys, accounts, and credentials belong directly to your practice.
[ ] **Staff Notification Verification:** Verify practice manager/receptionist receives real-time push alert with dial link.

FREQUENTLY ASKED QUESTIONS

Q: Is $0 really $0, or is there a catch?

A: The quotas are real and the demo fits inside them — but GCP requires billing enabled, so the catch is discipline, not a hidden fee. Set the budget alert, watch the three meters, and $0 holds. Exceed a quota and metered pricing begins immediately.

Q: Can the $0 agent answer my real business phone number?

A: No — and that's the line everyone blurs. A real phone number means PSTN, and PSTN is always metered. The $0 build is a browser-side agent on your website. The phone number is the paid upgrade you make after the concept is proven.

Q: What happens when I outgrow the free tier?

A: Two upgrade paths: pay GCP metered rates as usage grows (cheap at small scale), or move to a managed stack like Vapi + Twilio when you need a phone number, call recording compliance, and uptime guarantees. Both are visible line items — the recipe exists so you can choose with evidence.