Manual de referencia: Vapi

Datos de esta manual

Fase	MAN — Manuales de referencia
Módulo	MAN — Biblioteca de consulta

Objetivo práctico

Este manual reúne, en un único documento de consulta, la arquitectura completa del panel de Vapi (asistentes, modelo, voz, transcripción, herramientas, Webhooks, números, logs), sus planes y precios vigentes, la seguridad de sus claves de API, un examen honesto de la accesibilidad real de su panel con NVDA —incluidas las partes que presentan barreras— y el catálogo de errores frecuentes con su diagnóstico. Úsalo como referencia puntual, no como lectura lineal.

Resultado que obtendrás

Un documento de consulta permanente al que volver mientras trabajas con Vapi: desde la primera configuración de un asistente hasta el diagnóstico de un fallo concreto en producción meses después.

Dónde encaja dentro del sistema final

Dentro del sistema TELÉFONO → AGENTE → CEREBRO → VOZ → AUTOMATIZACIÓN → CALENDARIO, Vapi es una de las dos plataformas de referencia de este curso (junto a Retell) para la pieza AGENTE: la centralita que conecta STT (oído), LLM (cerebro) y TTS (voz) en una única conversación coherente, y que además expone la puerta de entrada TELÉFONO cuando se le asocia un número real. Sus «Tools» son también el punto de conexión hacia AUTOMATIZACIÓN (Webhooks hacia n8n o Make) y, por extensión, hacia CALENDARIO cuando una herramienta consulta o reserva disponibilidad.

Conocimientos previos

- Haber completado la lección f2-m3-l1-vapi.php (ensamblaje del agente en Vapi) dentro de la progresión del curso.
- Prompt Maestro completo (módulo 1.3).
- Concepto de Webhook y Tool Calling (módulos 1.2 y 3.4).
- Protocolo de creación segura de cuentas y de gestión de claves de API del módulo 2.1.

Herramientas necesarias

- Una cuenta de Vapi.
- Un navegador con NVDA activo.
- Un gestor de contraseñas para las claves de API.

Posible coste

Modelo de precios (verificado el 30 de agosto de 2026 contra la página oficial de precios, `vapi.ai/pricing`): Vapi factura por uso puro, sin planes de suscripción con funciones escalonadas. La tarifa base de orquestación es de 0,05 USD por minuto de llamada; a esa cifra se suma el coste real (a precio de coste, «at-cost») de los proveedores que elijas para transcripción (STT), modelo (LLM), voz (TTS) y telefonía. En la práctica, el coste total por minuto de un agente completo suele superar ampliamente los 0,05 USD anunciados como tarifa base: sumando todos los componentes, el rango habitual observado para una configuración completa se sitúa aproximadamente entre 0,10 y 0,30 USD por minuto, según los proveedores elegidos.

Concurrencia: el plan de autoservicio incluye por defecto 10 llamadas simultáneas; ampliar ese número tiene un coste adicional por línea y mes (aproximadamente 10 USD por línea adicional, verifica la cifra vigente en tu panel).

Mensajería: los mensajes de SMS/chat gestionados a través de Vapi tienen un coste aproximado de 0,005 USD por mensaje.

Crédito de prueba: Vapi ofrece crédito inicial para empezar a experimentar sin necesidad de introducir un método de pago de inmediato. La cantidad exacta vigente puede variar; compruébala en tu propio panel al registrarte.

Plan Enterprise: precio a medida, para volumen alto y requisitos adicionales de cumplimiento o soporte.

Límite de gasto nativo: según la documentación oficial de precios consultada, no se menciona ninguna función de límite de gasto configurable (hard limit) dentro de la propia plataforma. Esta situación coincide con lo ya verificado en la lección de límites de gasto del módulo 2.1: aplica medidas externas (saldo mínimo, tarjeta virtual con tope propio, alertas bancarias, monitorización activa) si vas a hacer pruebas o despliegues extensos.

Advertencia de seguridad

Dos tipos de clave de API, con alcance muy distinto: según la documentación oficial de Vapi, existen claves públicas, pensadas para código de lado del cliente (navegador), y claves privadas, pensadas exclusivamente para el lado del servidor. La clave privada nunca debe exponerse en el navegador: debe mantenerse en el servidor y leerse desde variables de entorno; no debe compartirse ni incluirse en ningún commit de control de versiones.

Restricción de origen de la clave pública: Vapi permite vincular una clave pública a una lista de orígenes permitidos (dominios y puertos concretos desde los que puede usarse). Si una petición de creación de llamada llega desde un dominio que no está en esa lista, Vapi la rechaza. Configura siempre esta restricción si vas a usar la clave pública en un widget de voz embebido en una web.

Autenticación de Webhooks: al configurar el servidor que recibe eventos y llamadas a herramientas desde Vapi (Server URL), la documentación oficial recomienda autenticar ese endpoint mediante credenciales personalizadas («Custom Credentials») referenciadas por un `credentialId`, en lugar de dejar la URL abierta sin verificación. Vapi permite además definir Server URLs a distintos niveles (herramienta, asistente, número de teléfono, cuenta), lo que te permite repartir la autenticación según la sensibilidad de cada integración.

Cifrado de argumentos de herramientas: para herramientas que manejan datos sensibles, Vapi ofrece cifrado de los argumentos mediante un par de claves pública/privada generado con OpenSSL, cuya clave pública se registra en el propio panel.

Autenticación JWT para clientes web: como alternativa más restrictiva a la clave pública simple, Vapi admite autenticación mediante JWT firmado con la clave privada en el servidor, para limitar más finamente lo que un cliente web concreto puede hacer.

Modelo mental

Piensa en Vapi como en una centralita telefónica completa con un telefonista integrado: no es el propio telefonista (el LLM), ni sus cuerdas vocales (el TTS), ni su oído (el STT) por separado, sino el mueble que los conecta a todos, gestiona la llamada de principio a fin y, además, tiene un cuadro de mandos con botones que puede pulsar durante la conversación (las Tools) para avisar a otros departamentos.

Explicación

Según la documentación oficial de Vapi verificada el 30 de agosto de 2026, un

Asistente es la unidad reutilizable de configuración: agrupa modelo (LLM), voz (TTS), transcriptor (STT), herramientas, prompts y comportamiento de llamada. Se crea desde «Dashboard → Assistants → Create Assistant», con un preajuste equilibrado por defecto («Balanced») que puede sustituirse eligiendo manualmente cada proveedor (OpenAI, Anthropic o Google para el modelo; ElevenLabs entre otros para la voz; Deepgram o Gladia entre otros para la transcripción).

Tools (herramientas): es la sección donde se definen las funciones que el LLM puede invocar durante la conversación. Cada herramienta se asocia normalmente a un Webhook: cuando el modelo decide usarla, Vapi envía una petición a la Server URL configurada, con el nombre de la función y sus parámetros, y espera una respuesta para continuar la conversación con esa información. Este es el mecanismo base del Tool Calling trabajado en el módulo 3.4.

Server URLs y eventos: Vapi permite configurar una URL de servidor a distintos niveles (por herramienta, por asistente, por número de teléfono o a nivel de cuenta), y notifica una variedad de eventos del ciclo de vida de la llamada (inicio, fin, actualización de estado, mensajes de función, entre otros) mediante ese mismo mecanismo de Webhook.

Workflows (flujos): además del modo «Asistente» (conversación libre guiada por un único prompt, el enfoque usado en este curso), Vapi ofrece «Workflows», un editor visual de nodos y conexiones para diseñar flujos de conversación deterministas: nodos de conversación (diálogo con prompt propio), nodos de petición a una API externa, nodos de finalización de llamada, y bloques como «Say» (respuesta fija o dinámica), «Gather» (recogida de datos del usuario) y «Condition» (ramificación de la conversación). Los Workflows se construyen arrastrando bloques sobre un lienzo y trazando conexiones entre ellos.

Números de teléfono: un número de teléfono en Vapi puede tener un asistente asignado de forma fija, o puede decidirse dinámicamente en el momento de la llamada mediante una petición al servidor configurado, si el número no tiene asistente fijo.

Grabación, transcripción y logs: las grabaciones, transcripciones y registros de cada llamada están disponibles tanto a través de la API como en el propio panel del Dashboard, y pueden activarse o ajustarse mediante el «artifactPlan» del asistente.

Vocabulario nuevo

Assistant (Vapi)

Piénsalo así: El telefonista completo ya montado: cerebro, oído y voz configurados juntos.

Configuración reutilizable de un agente en Vapi: modelo, voz, transcriptor, prompt, primer mensaje y herramientas asociadas.

Ejemplo: El asistente «Recepción Fontanería Ejemplo» construido en la Fase 6 del curso.

Server URL

Piénsalo así: La dirección del buzón al que Vapi envía cada aviso importante durante la llamada.

URL de tu propio servidor a la que Vapi envía peticiones HTTP (Webhooks) para invocar herramientas o notificar eventos de la llamada. Puede configurarse a nivel de herramienta, asistente, número o cuenta.

Ejemplo: Una Server URL apuntando a un Webhook de n8n que registra cada cita reservada por el agente.

Workflow (Vapi)

Piénsalo así: Un diagrama de flujo dibujado a mano, con cajas y flechas, en lugar de un guion escrito en prosa.

Editor visual de nodos y conexiones de Vapi para diseñar conversaciones deterministas paso a paso, como alternativa al modo «Asistente» de prompt único.

Ejemplo: Un flujo de cualificación de leads con nodos de pregunta, condición y transferencia según la respuesta.

Clave pública / clave privada (Vapi)

Piénsalo así: La llave del buzón de recepción, que puede usar cualquier visitante, frente a la llave maestra del edificio, que solo debe tener el conserje.

La clave pública está pensada para código de cliente (navegador) y puede restringirse por dominio de origen. La clave privada es exclusiva del servidor y nunca debe exponerse en el navegador.

Ejemplo: Un widget de llamada de voz embebido en una web pública usa la clave pública, restringida a ese dominio.

artifactPlan

Piénsalo así: La orden de qué documentar de cada llamada: grabación, transcripción, ambas o ninguna.

Configuración del asistente que determina qué artefactos (grabación de audio, transcripción, registro detallado) se generan y conservan para cada llamada.

Ejemplo: Activar `loggingEnabled` en el artifactPlan para obtener registros detallados de cada llamada de un cliente.

Preparación

Ten a mano tu Prompt Maestro, la clave de API de ElevenLabs u otro proveedor de voz, y decide de antemano si vas a usar el modo Asistente (recomendado por este curso) o Workflows antes de empezar a construir, porque cambiar de modo a mitad de proceso obliga a reconstruir la configuración desde cero.

Procedimiento paso a paso

1. **Contexto:** Vas a generar una clave de API privada para conectar tu propio backend con Vapi.

   Acción de teclado: Recorre encabezados con `H` hasta la sección de claves de API del panel (suele estar bajo «Settings» o el nombre de tu organización). Activa «Create Key» o equivalente con `Enter`.

   Respuesta esperada de NVDA: NVDA debería anunciar un texto similar a «Create Key, botón» y, tras activarlo, un campo de nombre y un selector de tipo de clave (pública o privada).

   Qué significa: Elige «privada» si el uso será exclusivamente desde tu propio servidor; elige «pública» solo si el código se ejecutará en un navegador y vas a restringir el origen.

   Acción siguiente: Guarda la clave generada en tu gestor de contraseñas de inmediato: no volverá a mostrarse completa después de este paso.

2. **Contexto:** Vas a configurar una Tool (herramienta) conectada a un Webhook externo.

Acción de teclado: Dentro del editor de tu asistente, recorre encabezados con `H` hasta «Tools». Activa «Add Tool» o «Create Tool» con `Enter`, completa con `Tab` el nombre de la función, su descripción, los parámetros esperados y la Server URL de destino.

Respuesta esperada de NVDA: NVDA anunciará cada campo del formulario conforme lo recorras, incluidos los campos de tipo (texto, número, booleano) de cada parámetro.

Qué significa: La descripción de la función es lo que el LLM lee para decidir cuándo invocarla: una descripción vaga produce invocaciones erráticas.

Acción siguiente: Guarda la herramienta, asígnala al asistente si no se ha vinculado automáticamente, y pruébala mediante una llamada web antes de conectarla a un número real.

3. **Contexto:** Vas a revisar los logs de una llamada ya finalizada.

   Acción de teclado: Recorre encabezados con `H` hasta «Call Logs» o «Calls». Localiza la llamada en la tabla de resultados con `Tab` o con la navegación por tablas de NVDA (`Ctrl` + `Alt` + flechas), y actívala con `Enter` para abrir el detalle.

   Respuesta esperada de NVDA: NVDA debería anunciar las columnas de la tabla (fecha, duración, estado, asistente) y, al abrir el detalle, la transcripción y los eventos registrados.

   Qué significa: El detalle de la llamada es la fuente más fiable para diagnosticar un comportamiento inesperado del agente, más que confiar en la memoria de la prueba.

   Acción siguiente: Descarga la grabación o la transcripción si necesitas revisarla fuera del panel, o compártela con el equipo si vas a depurar un error concreto.

Posibles diferencias de interfaz

Vapi actualiza su panel con cierta frecuencia. Los conceptos (asistente, modelo/voz/transcriptor, herramientas, Server URL, números, logs) son estables aunque el nombre exacto de un botón o la disposición de un menú cambien entre versiones.

Problema de accesibilidad y alternativa

Barrera real identificada: el editor visual de «Workflows» es un lienzo de nodos y conexiones que se construye principalmente arrastrando bloques con el ratón y trazando líneas entre puntos de conexión de cada nodo (drag and drop). Esta interacción no tiene, según la documentación pública consultada, un equivalente declarado de teclado o de lector de pantalla dentro del propio editor visual: mover un nodo, conectarlo a otro y ajustar su posición en el lienzo depende de una interacción de arrastre que NVDA no puede interpretar de forma fiable sobre un canvas gráfico. Esto constituye una barrera de accesibilidad real para construir flujos complejos exclusivamente desde el panel visual con NVDA.

Alternativa disponible: tanto los Asistentes como los propios Workflows pueden crearse y modificarse por completo mediante la API REST de Vapi (los endpoints de creación de asistentes y de flujos aceptan una definición completa en JSON, incluida la disposición lógica de nodos y conexiones de un Workflow). Esta es la vía recomendada para construir cualquier lógica compleja con NVDA: define el flujo como una estructura de datos en tu editor de código habitual (accesible, con autocompletado y sin arrastre), y envíalo a Vapi mediante la API, en lugar de operar el lienzo visual directamente. El modo «Asistente» de prompt único —el que usa este curso— se configura íntegramente mediante formularios estándar (campos de texto, selectores, listas), y no presenta esta barrera: es accesible con NVDA sin necesidad de recurrir a la API.

Errores frecuentes y diagnóstico

Una herramienta (Tool) nunca recibe la petición esperada en el servidor.

Cómo localizarlo: Revisa que la Server URL configurada sea correcta y esté accesible públicamente, que la descripción de la función sea lo bastante clara para que el LLM decida invocarla, y consulta el log detallado de la llamada para confirmar si Vapi intentó la petición y qué respuesta obtuvo.

Solución: Corrige la URL o la descripción de la función, y usa una herramienta de inspección de peticiones (como un endpoint de prueba temporal) para confirmar que la petición llega con el formato esperado.

La llamada de prueba por voz desde el navegador no capta el micrófono.

Cómo localizarlo: Puede deberse a permisos de micrófono no concedidos al navegador para el dominio de Vapi, o a un micrófono no seleccionado como predeterminado en el sistema operativo.

Solución: Revisa los permisos de micrófono del navegador para el sitio de Vapi, y confirma en la configuración de sonido de Windows qué dispositivo está activo por defecto.

El coste real de las llamadas es notablemente superior a la tarifa de 0,05 USD/minuto anunciada como base.

Cómo localizarlo: La tarifa de 0,05 USD/minuto es solo la orquestación de Vapi; a ella se suman, a precio de coste, los proveedores de STT, LLM, TTS y telefonía elegidos, que pueden multiplicar el coste total por minuto varias veces.

Solución: Antes de presupuestar un proyecto, calcula el coste combinado real sumando la tarifa de Vapi y el precio vigente de cada proveedor elegido, no solo la cifra publicitada de la orquestación.

El asistente responde en un idioma distinto al esperado.

Cómo localizarlo: El transcriptor (STT) o el modelo (LLM) pueden no estar configurados explícitamente en español, o el Prompt Maestro no incluye una instrucción explícita de idioma.

Solución: Revisa la configuración del transcriptor y añade una instrucción explícita de idioma en el prompt del sistema si es necesario.

Comprobación final

Confirma que puedes, usando solo NVDA: crear una Tool con su Server URL, localizar el detalle de logs de una llamada de prueba anterior, y explicar por qué el modo Asistente es preferible a Workflows cuando trabajas exclusivamente desde el panel visual. Si consigues las tres cosas, la comprobación es positiva.

Qué acabas de conseguir

Tienes una referencia completa de la arquitectura de Vapi —asistentes, herramientas, Webhooks, números y logs—, con un criterio claro sobre qué partes del panel son accesibles de forma directa con NVDA y cuáles requieren usar la API como alternativa.

Ejercicio práctico

Diseña, solo sobre el papel o en un editor de texto, la definición JSON de un Workflow sencillo de dos nodos (una pregunta y una condición), como ejercicio de cómo abordarías su creación por API en lugar de por el lienzo visual.

Resumen de lo imprescindible

- Vapi cobra 0,05 USD/minuto de orquestación más el coste real de los proveedores de STT, LLM, TTS y telefonía elegidos: el coste total suele ser varias veces superior a esa tarifa base.
- La documentación oficial de precios de Vapi no menciona un límite de gasto configurable nativo: aplica medidas externas de control.
- La clave privada es exclusiva de servidor; la clave pública puede restringirse por dominio de origen para uso en cliente.
- El editor visual de Workflows (drag and drop de nodos) es una barrera real de accesibilidad con NVDA; la API REST permite construir la misma lógica sin depender del lienzo.
- El modo Asistente, usado por este curso, es completamente accesible con NVDA mediante formularios estándar.

Estoy preparado para continuar si puedo…

- Explicar la diferencia entre el modo Asistente y Workflows en Vapi, y cuál es más adecuado con NVDA.
- Configurar una Tool con su Server URL y diagnosticar por qué no llega una petición esperada.
- Calcular el coste real por minuto de una configuración de Vapi, más allá de la tarifa base publicitada.
- Proteger correctamente una clave privada y restringir una clave pública por dominio de origen.

Fuentes

- Vapi Pricing — Vapi (consultado el 2026-08-30)
- Assistants quickstart — Vapi (consultado el 2026-08-30)
- Introduction to Tools — Vapi (consultado el 2026-08-30)
- Server URLs — Vapi (consultado el 2026-08-30)
- Server authentication — Vapi (consultado el 2026-08-30)
- Workflows overview — Vapi (consultado el 2026-08-30)
- Proxy server guide (seguridad y privacidad) — Vapi (consultado el 2026-08-30)
- Provider Keys — Vapi (consultado el 2026-08-30)
- Call recording, logging and transcribing — Vapi (consultado el 2026-08-30)

Última actualización de este contenido: **2026-08-30**.