FICHA TÉCNICA DE PRODUCTO

VoiceAI Gateway

IVR Conversacional Inteligente y Enrutamiento Semántico para Contact Centers

Atributo	Detalle
Categoría	Telefonía conversacional con IA / IVR inteligente
Plataforma objetivo	Contact Centers sobre Asterisk 16+ y Vicidial 2.14+
Despliegue	On-Premise o Nube Privada; procesamiento 100% local
Modelos de IA	Modelos locales de código abierto: reconocimiento de voz, **modelo decisor** (LLM de clasificación) y síntesis de voz neuronal
Integración	FastAGI (TCP :4573) y despachador de In-Groups de Vicidial
Idioma	Español neutro (taxonomía y prompts configurables)

01 Resumen Ejecutivo

VoiceAI Gateway es una solución integral de telefonía conversacional con Inteligencia Artificial, desplegada On-Premise o en Nube Privada, que sustituye los árboles de decisión DTMF tradicionales (menús de *"marque 1 para..."*) por un asistente de voz fluido y natural.

El sistema escucha al interlocutor en lenguaje natural, transcribe su mensaje en tiempo real y lo entrega a un **modelo decisor local**: un modelo de lenguaje de código abierto ejecutado dentro de la infraestructura del cliente que interpreta la intención, resuelve ambigüedades y decide hacia qué cola de agentes (In-Group / skill-based routing) o flujo de autoservicio debe enrutarse la llamada, de forma transparente, sobre plataformas de Contact Center basadas en Asterisk / Vicidial.

02 Diagrama de Arquitectura de Alto Nivel

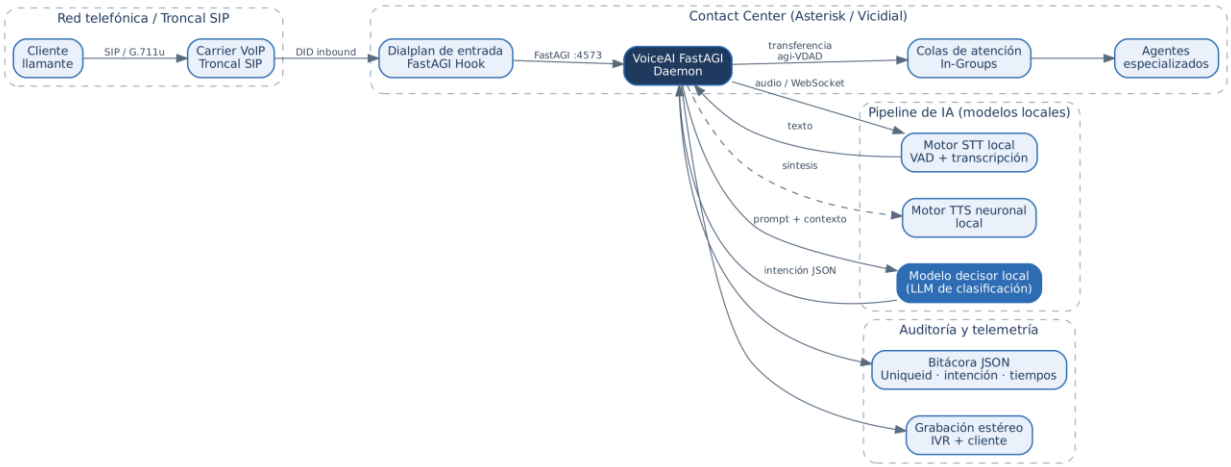


Figura 1. Flujo de la llamada entre la red telefónica, la PBX, el pipeline de modelos locales y la capa de auditoría.

03 Componentes y Especificaciones Técnicas

Módulo	Tecnología / Protocolo	Función principal
Controlador de Telefonía	FastAGI (Asterisk Gateway Interface) sobre TCP asíncrono	Orquesta el ciclo de vida de la llamada (Answer, Playback, Record, Transfer, Hangup) con latencia inferior a 50 ms.
PBX / ACD Host	Asterisk 16+ / Vicidial Core	Señalización SIP, troncales E1/SIP, DIDs entrantes y distribución automática de llamadas (ACD).
Motor de Reconocimiento (STT)	Modelo local de voz a texto vía WebSocket streaming	Transcripción en tiempo real con detección de actividad de voz (VAD) adaptada a audio telefónico (PCM 8 kHz / 16 bits).
Modelo Decisor (NLU / LLM)	Modelo de lenguaje local de código abierto sobre servidor de inferencia local (8B a 31B parámetros)	Clasificación de intención zero-shot, desambiguación semántica, extracción de entidades y decisión de enrutamiento en JSON estricto.
Motor de Voz (TTS neuronal)	Modelo local de síntesis de voz neuronal con clonación de voz	Voz natural hiperrealista, normalización fonética y adaptación a retardos de jitter buffer RTP.
Enrutador de Colas	Vicidial In-Group Dispatcher (`agi-VDAD_ALL_inbound.agi`)	Inyecta la llamada directamente en el grupo de habilidades decidido, sin intervención manual.
Auditoría y Grabación	Ensamblado estéreo y extracción directa de audio	Captura íntegra: audio del IVR, voz del cliente y bitácora con transcripción, decisión y métricas de inferencia.

04 Modelo Decisor Local

El núcleo de inteligencia del producto es un **modelo de lenguaje local de código abierto** que actúa como motor de decisión. A diferencia de un IVR por palabras clave, el modelo decisor comprende la petición completa del cliente, tal como la expresa, y determina el destino óptimo de la llamada.

Responsabilidades

- **Comprensión de intención:** interpreta frases libres, coloquiales o incompletas sin requerir comandos de voz rígidos.
- **Desambiguación:** cuando la intención no es clara, genera una pregunta de aclaración breve antes de decidir.
- **Extracción de entidades:** identifica datos relevantes (número de cuenta, tipo de servicio, motivo) para adjuntarlos al contexto de la transferencia.
- **Decisión de enrutamiento:** selecciona la categoría de negocio y el In-Group destino a partir de la taxonomía configurada.

Entrada y salida

Elemento	Descripción
Entrada	Transcripción del cliente + prompt de sistema con la taxonomía vigente + contexto de la llamada (DID, horario, turno de diálogo).
Salida	Objeto JSON estricto con `intencion`, `categoria`, `in_group`, `confianza`, `entidades` y `requiere_aclaracion`.
Política de decisión	Si la confianza supera el umbral configurado, se transfiere; si no, se solicita aclaración; tras reintentos fallidos se deriva a Atención General.
Ejecución	Inferencia local en GPU dentro de la red del cliente, sin llamadas a APIs externas.

Ventajas de un decisor local

- Independencia de proveedores externos y costos por consumo de API.
- Los datos de voz y texto nunca salen de la infraestructura del cliente.
- Modelo intercambiable: puede actualizarse o ajustarse a la taxonomía del negocio sin rediseñar la arquitectura.

05 Flujo Operativo de la Llamada

1. El cliente marca al DID; la troncal SIP entrega la llamada a Asterisk.
2. El dialplan invoca al **VoiceAI FastAGI Daemon**, que contesta, inyecta 500 ms de micro-silencio y reproduce el saludo y el aviso de privacidad.
3. El audio del cliente se envía al **motor STT local**, que detecta fin de habla y devuelve la transcripción.
4. El **modelo decisor local** analiza la transcripción y devuelve la decisión en JSON.
5. El daemon confirma al cliente mediante el **motor TTS local** y transfiere la llamada al In-Group decidido.
6. Se cierran la grabación estéreo y la bitácora de metadatos (Uniqueid, intención, decisión, tiempos).

06 Categorías de Enrutamiento (Skill-Based Routing)

El modelo decisor opera sobre una taxonomía modular configurable. Por defecto incluye las siguientes áreas de negocio:

Categoría	Intenciones detectadas
Retención / Cancelaciones	Bajas de servicio, inconformidades graves, quejas recurrentes o solicitudes de cancelación.
Soporte Técnico	Fallas de conectividad, equipos dañados o interrupciones de servicio.
Servicios y Facturación	Estado de cuenta, facturación, fechas de corte, aclaración de cobros y saldos.
Ventas y Contrataciones	Nuevos planes, adición de servicios, promociones y consultas de cobertura.
Cobranza y Pagos	Reporte de pagos, reactivación de servicios y promesas de pago.
Atención General / Humano	Solicitud explícita de operador o consultas fuera de catálogo.

07 Tolerancia a Fallos y Experiencia de Usuario

- **Manejo de silencios y reprompting:** si el cliente no responde tras el saludo, se emite un recordatorio de cortesía (*"Disculpe, ¿sigue en la línea?"*); si el silencio persiste, el sistema se despide y libera el canal.
- **Fail-Safe Transfer:** ante indisponibilidad de los modelos locales o fallas de transcripción, la llamada nunca se corta; se transfiere de inmediato a la cola general con un agente humano.
- **Normalización fonética:** preprocesamiento de URLs, dominios, abreviaturas y dígitos para una pronunciación natural en español neutro.
- **Compensación de apertura RTP:** micro-silencio de 500 ms al inicio para que la primera sílaba no se recorte durante la negociación de codecs SIP.

08 Métricas y Rendimiento (KPIs)

Indicador	Valor objetivo
Tiempo total de decisión	< 1.2 s desde que el cliente termina de hablar hasta la confirmación y transferencia
Precisión de clasificación semántica	> 95% en pruebas de lenguaje natural
Formato de grabación	WAV / PCM 8,000 Hz, 16 bits, mono o dual-channel
Disponibilidad	99.9% con microservicios desacoplados
Latencia de control de llamada	< 50 ms por comando FastAGI

09 Seguridad, Cumplimiento y Privacidad

- **Procesamiento 100% local:** audio, transcripción y decisión se ejecutan con modelos locales dentro de la red corporativa, sin envío de datos a APIs de terceros.
- **Aviso de privacidad configurable** al inicio de la llamada, conforme a la normativa vigente de protección de datos personales.
- **Comunicaciones aisladas:** canales internos sobre LAN/VLAN segregadas con autenticación centralizada.
- **Trazabilidad completa:** cada decisión del modelo queda registrada con su transcripción y nivel de confianza para auditoría.

10 Requisitos de Infraestructura Recomendada

Componente	Especificación
Servidor PBX	Asterisk 16+, Vicidial 2.14+, red dedicada con ancho de banda SIP dimensionado
GPU (inferencia de modelos locales)	NVIDIA RTX 3090 / RTX 4090 (24 GB VRAM) o NVIDIA A10 / A30 / A100
CPU	8+ núcleos x86_64
RAM	32 GB o más
Almacenamiento	NVMe SSD para carga rápida de pesos de modelos
Sistema operativo	Ubuntu 22.04 LTS / Debian 12 / Rocky Linux 9, o despliegue en contenedores Docker

11 Beneficios Clave

- Elimina los menús DTMF y reduce la fricción del cliente desde el primer segundo de la llamada.
- Enrutamiento más preciso hacia el agente especializado, lo que disminuye transferencias internas y rellamadas.
- Soberanía de datos y costos predecibles gracias a modelos locales sin dependencia de nube pública.
- Integración no invasiva con la operación Vicidial existente: colas, agentes y reportes se mantienen.
- Grabaciones y bitácoras listas para speech analytics, calidad y cumplimiento.