

Como construir um voice agent

Do ramal no PABX ao webhook de fechamento: o caminho da mídia, o orçamento de latência e as topologias de conexão para a plataforma feita para voz.

Base da API <https://api.sippulse.ai/v1> Autenticação `api-key: $SIPPULSE_API_KEY`
Proxy SIP `sip.sippulse.ai:5060` • Protocolo UDP / TCP • URI do agente `{agent_id}@sip.sippulse.ai`

1 Arquitetura em tempo real: caminho da mídia e orçamento de latência

SINALIZAÇÃO SIP — MÍDIA RTP & PIPELINE DE STREAMING



ORÇAMENTO DE LATÊNCIA — ALVO DE 800 MS BOCA A OUVIDO

50 ms	150 ms	400 ms	150 ms	50 ms
VAD	STT streaming	LLM (TTFT)	1ª chunk TTS	Rede

- O LLM consome metade do orçamento:** otimize TTFT com prompts curtos, poucas ferramentas no contexto e modelos rápidos.
- Inicie o TTS na primeira frase pronta:** o áudio começa a ser transmitido antes de o LLM concluir toda a resposta.
- Regra prática:** pausa acima de 1 s é percebida pelo usuário como falha na ligação — 200 ms é imperceptível, 500 ms já soa lento.

2 Integração: como o agente entra na infraestrutura

- Ramal no seu PABX (melhor posição):** o agente ocupa a posição de um ramal. O PABX continua dono do plano de discagem — fila, URA, transbordo e gravação seguem valendo.
- A SipPulse registra o agente no seu PABX:** External SIP Registration faz o agente se registrar como ramal usando `Servidor SIP`, `Usuário` e `Senha`. Beta, habilitado por organização.
- URI SIP direta:** `{agent_id}@sip.sippulse.ai` — o PABX encaminha a chamada para a URI, sem precisar comprar número.
- Tronco SIP:** `sip.sippulse.ai:5060` (UDP/TCP). Reserve para multi-tenant, alto volume e separação de redes.
- DID nacional:** adquira na aba Números (exclusivo Brasil), um DID por agente.

CONTEXTO ANTES DO “ÁUDIO” — SIP HEADERS (BASE64)

- `X-Additional-Instructions`: concatena contexto dinâmico às instruções do agente.
- `X-Initial-Message`: substitui a mensagem inicial do agente.
- `X-Uid`: ID único de sessão para correlação direta no CRM ou ticket.

3 Outbound: duas formas de originar chamadas

- A. Seu discador transfere a chamada atendida:** conecta no PABX (fila, URA, lista) e transfere para o ramal ou a URI do agente com os headers em Base64.
- B. API de discagem outbound da SipPulse AI:**

```
POST /agents/{id}/outbound-call base: api.sippulse.ai/v1
{
  "phone_number": "+5511999999999",
  "metadata_overrides": {
    "initial_message": "Olá, posso te ajudar?",
    "additional_instructions": "Fatura 4/7, 12 dias em atraso.",
    "variables": { "parcela": 3, "vencimento": "12/09" }
  }
}
```

- De qual número a chamada sai:** alugue o DID na SipPulse (por DDD) ou traga o seu.
- Tools de voz não entram em `metadata_overrides`:** ficam no agente e valem para todas as chamadas.

Regra prática: o outbound só supera o inbound quando `additional_instructions` traz o motivo daquela ligação.

4 Roteamento: onde o agente entra no fluxo

- Se o seu PABX já sabe rotear atendimento humano, ele já sabe rotear o agente — é só mais um ramal.
- Tráfego no lugar da URA:** entende o motivo em linguagem natural e entrega a chamada qualificada.
- Transbordo de fila:** atende automaticamente quando a espera estoura o limite.
- Fora do horário:** cobertura 24/7 com registro e triagem de chamados.
- Ramal dedicado:** ramal exclusivo para fluxos de segurança, cobrança ou agendamento.

SAÍDA PARA O ATENDIMENTO HUMANO (HANDOFF)

- `transfer_call` via SIP REFER devolve a chamada para ramal, fila ou URA.
- Passe contexto junto:** transmita os dados coletados na transferência para evitar repetições pelo cliente.
- Gatilhos claros:** defina transbordo por silêncio, palavra-chave ou intenção fora de escopo.

5 Configuração do agente

- Ferramenta principal:** voz (telephony, 8 kHz, mono).
- Modelo:** escolha entre velocidade e precisão (TTFT x qualidade).
- Tools e MCP:** conecte APIs e sistemas para consulta e ação.
- Parâmetros do LLM:** `temperature`, `top_p`, `max_tokens`, `reasoning_effort`.
- Voz e idioma:** seleção de voz TTS e idioma (controla STT e TTS).
- Sons de espera:** áudios AI Generated, Smart ou fixos.
- Limites de chamada:** duração de 0,5 a 60 min; inatividade de 1 a 300 s.
- Pós-análise:** análises estruturadas ao fechar a chamada.

6 Prompting para voz

- Regras de conversa:** tom, abertura direta, escopo claro e solução de ambiguidades.
- Estilo de saída:** respostas curtas e concisas. Nunca cole o histórico da conversa.
- Ferramentas e limites:** defina com precisão até onde o agente pode atuar.
- Prova de ação:** confirme status, números de protocolo e dados cadastrais.
- Timing de fala:** dê espaço para pausas naturais do usuário sem interromper precocemente.
- Saudação:** AI Generated, Smart (contextual) ou mensagem fixa.

7 Tools: agente virando sistema

TOOL DE VOZ	ONDE FUNCIONA
end_dialog	TODOS OS CANAIS
transfer_call	SÓ SIP
send_dtmf	SÓ SIP
receive_dtmf	SÓ SIP (BETA)
SIP REFER: transferência para URI/ramal (<code>sip:suporte@acme.com</code>), timeout de 1 a 15 s — padrão 5 s. DTMF de entrada: timeout de 1 a 60 s. Releia os dígitos e confirme antes de agir. Playground: as tools de voz só funcionam em chamadas SIP reais. <code>end_dialog</code> devolve JSON Schema com desfecho e motivo.	

8 Webhooks e análise

EVENTOS-CHAVE DE VOZ
<code>voice.start_call</code> → conexão estabelecida <code>voice.end_call</code> → encerramento da mídia <code>thread.closed</code> → sessão finalizada <code>structured_analysis.completed</code> → pós-análise pronta <code>credit.low_balance</code> → alerta de saldo
SEGURANÇA & ASSINATURA
- HMAC SHA256 (hex) sobre <code>{payload}:{timestamp}</code>, entregue em <code>x-signature</code> e <code>x-timestamp</code>. - Rejeite timestamps fora da janela de ±5 min contra ataques de replay. - Entrega com até 3 novas tentativas em caso de falha.

9 Pós-chamada & checklist

PÓS-CHAMADA (CRM & AUDITORIA)
- Resumo estruturado, intenção e taxa de desfecho. - Score de confiança do agente. - Injeção direta no CRM via webhook/API.
CHECKLIST DE PRÉ-PRODUÇÃO
Registro do ramal: agente registrado e discável a partir do PABX. Mídia OK: RTP, jitter e zero perda de pacote. Headers SIP: validação da decodificação Base64. Handoff & DTMF: teste de transbordo e dígitos. Barge-in: verificação de interrupção imediata.