Integração de Agentes de Inteligência Artificial com Sistema de Voz Convencional via VoIP e Asterisk

Arthur Henrique Tavares de Lyra Costa[1,2], **Renato Mariz de Moraes**[1]

[1]Centro de Informática (CIn) – Universidade Federal de Pernambuco (UFPE)
[2]Loomi – Porto Digital
Recife – Pernambuco – Brasil

arthur@loomi.com.br, renatomdm@cin.ufpe.br

Resumo. *A modernização de sistemas legados de atendimento por voz é um desafio relevante para organizações que dependem de infraestruturas TDM e não podem substituir integralmente seus equipamentos. Este trabalho propõe e avalia uma arquitetura híbrida que integra um PABX tradicional a agentes de Inteligência Artificial na nuvem por meio de VoIP e Asterisk. A solução implementa um pipeline conversacional completo, envolvendo reconhecimento automático de fala (ASR), modelo de linguagem (LLM) e síntese de voz (TTS), permitindo substituir menus DTMF por uma interação natural. Um protótipo foi validado em cenário real, com enlace digital E1, e submetido a testes de desempenho. Os resultados indicaram alta precisão de transcrição (WER de 2,65%) e latência média de 1.438 ms. A análise mostrou que uma parcela significativa do atraso decorre da infraestrutura de teste, e não do processamento de IA. Conclui-se que a abordagem híbrida é tecnicamente viável e permite modernizar sistemas legados sem a substituição completa do hardware.*

Abstract. *Modernizing legacy voice-based customer service systems remains a significant challenge for organizations that rely on TDM infrastructures and cannot fully replace existing hardware. This work proposes and evaluates a hybrid architecture that integrates a traditional PABX with cloud-based Artificial Intelligence agents through VoIP and Asterisk. The solution implements a complete conversational pipeline, including Automatic Speech Recognition (ASR), a Large Language Model (LLM), and Text-to-Speech (TTS), enabling natural voice interaction to replace rigid DTMF menus. A functional prototype was validated in a real-world scenario using a digital E1 link and subjected to performance testing. The results demonstrated high transcription accuracy (2.65% Word Error Rate) and an average end-to-end latency of 1,438 ms. The analysis showed that a significant portion of the delay was attributable to the testing infrastructure rather than to AI processing. The findings indicate that the proposed hybrid approach is technically feasible and enables modernization of legacy systems without complete hardware replacement.*

1. Introdução

No cenário atual de atendimento ao cliente, a eficiência e a qualidade da interação são fatores críticos para o sucesso empresarial. Contudo, muitas organizações ainda dependem de sistemas de telefonia legados, baseados na interação humana, que frequentemente resultam em longos tempos de espera e elevados custos operacionais. Embora a automação

por meio de *chatbots* tenha se tornado comum, a integração de agentes de Inteligência Artificial (IA) com canais de voz em tempo real permanece um desafio significativo, especialmente para empresas que não podem ou não desejam substituir completamente suas infraestruturas de telecomunicações existentes.

O presente trabalho busca modernizar os sistemas de telefonia convencionais para permitir um atendimento automatizado e inteligente, sem a necessidade de substituir, de forma dispendiosa, a infraestrutura legada. A integração da voz, que envolve o processamento de áudio em tempo real, apresenta complexidade adicional em relação à integração baseada em texto, exigindo uma arquitetura capaz de gerenciar a comunicação de forma fluida e com baixa latência. O objetivo geral deste trabalho é propor e validar esta arquitetura de sistema, capaz de integrar agentes de IA à infraestrutura de telefonia tradicional, visando à automação e à melhoria da qualidade do atendimento por voz.

O restante deste trabalho segue para a Seção 2, que analisa os trabalhos relacionados a esta proposta. A Seção 3 estabelece a fundamentação sobre VoIP, Asterisk e conexão com a IA conversacional. Na Seção 4, detalham-se a arquitetura da solução e a metodologia de implementação. Os resultados experimentais do protótipo são apresentados e discutidos na Seção 5. Por fim, a Seção 6 sintetiza as conclusões deste estudo e aponta direções para trabalhos futuros.

2. Trabalhos Relacionados

A literatura define o Asterisk como a plataforma padrão para implementação de PABX baseados em software [Madsen et al. 2013]. Estudos recentes reforçam a eficácia da implementação de *SIP Trunking* (tecnologia que substitui linhas telefônicas físicas por uma conexão via internet) em sistemas baseados em Asterisk como método para conectar redes IP a infraestruturas de telefonia, garantindo interoperabilidade e reduzindo custos [Cunha et al. 2012]. Esta abordagem é corroborada pela documentação técnica da comunidade, que detalha o uso do Asterisk como ponte para sistemas legados [VoIP Info 2025]. No contexto específico deste trabalho, a modernização foca na integração com o PABX IS 3000, um sistema TDM cuja arquitetura e limitações de conectividade IP são descritas em seus manuais técnicos de operação [Unify Software and Solutions 2020], o que justifica a necessidade de um *gateway* intermediário.

Por outro lado, estudos recentes propõem arquiteturas de PABX VoIP automatizadas baseadas em IA, especificamente para pequenas e médias empresas (SMEs), demonstrando como assistentes de voz podem substituir interações humanas em tarefas de roteamento [Fernando 2025]. Do ponto de vista da implementação técnica, projetos de código aberto demonstram a viabilidade do uso da interface EAGI (*Enhanced Asterisk Gateway Interface* ou Interface Melhorada de Portas Asterisk) para conectar o fluxo de áudio do Asterisk a serviços de reconhecimento de fala, como o Google Speech-to-Text [Sultan 2018]. O mercado também aponta para uma evolução nessa integração, com plataformas de síntese de voz, como a ElevenLabs, passando a oferecer integrações nativas de *SIP Trunking* para seus agentes conversacionais, sinalizando uma tendência de conexão direta entre a telefonia e os modelos generativos [ElevenLabs 2025].

No que se refere aos componentes do *pipeline* de IA (ASR, LLM e TTS), o estado da arte em modelos de diálogo falado (*Spoken Dialogue Models*) vem evoluindo rapidamente, com pesquisas que propõem arquiteturas em cascata e ponta-a-ponta

[Chen et al. 2024]. Para o reconhecimento de fala (ASR), *benchmarks* atuais comparam a precisão, a velocidade e o custo de diversas APIs, fornecendo dados essenciais para a escolha do fornecedor em aplicações de tempo real [VoiceWriter 2025], [McGillivray 2024]. Da mesma forma, a tecnologia de síntese de voz (TTS) é avaliada em estudos de revisão que comparam diferentes abordagens de geração [Chowdhury and Hussan 2023] e propõem novas métricas de distribuição para avaliar a qualidade e naturalidade da fala sintética [Minixhofer et al. 2024], [Hugging Face 2025]. No núcleo do processamento, os Grandes Modelos de Linguagem (LLMs) são continuamente avaliados em *leaderboards* que medem sua capacidade de resposta e raciocínio em diversos domínios [González-Bustamante 2024].

Estudos recentes sobre sistemas de avatares falados e turnos conversacionais destacam que a latência média humana em diálogos é de aproximadamente 239 ms na língua inglesa [Jacoby et al. 2024]. A literatura aponta que as arquiteturas atuais de IA têm dificuldade em atingir essa marca, criando lacunas de silêncio que comprometem a naturalidade da interação [Jacoby et al. 2024]. Este trabalho se posiciona justamente nesta lacuna, utilizando uma ferramenta de *softphone* consolidada [MicroSIP nd] para validar uma arquitetura híbrida e avaliar se a integração de sistemas legados com APIs de nuvem modernas opera dentro dos limites de latência aceitáveis definidos pelos estudos de interação humano-computador.

3. Fundamentação

Este trabalho envolve a interseção de quatro domínios tecnológicos distintos, cuja compreensão é essencial para justificar a arquitetura proposta. A solução opera traduzindo (1) a telefonia legada TDM do PABX IS 3000 para (2) os protocolos de Voz sobre IP (VoIP), em que (3) o Asterisk atua estritamente como um *gateway* de mídia e sinalização, permitindo que o sistema legado se conecte a (4) um agente de Inteligência Artificial externo via troncos SIP.

3.1. Telefonia Legada

A telefonia tradicional, representada neste trabalho pelo PABX IS 3000, opera sob o paradigma da *comutação de circuitos* (*circuit switching*). Por outro lado, o TDM é a tecnologia base da telefonia digital legada. Em um sistema TDM, um meio de transmissão físico de alta capacidade é dividido em múltiplos intervalos de tempo (*time slots*) fixos e recorrentes.

A rede garante que, para cada chamada, um circuito físico ou lógico de ponta a ponta é estabelecido. Durante a chamada, um par de *time slots* (um para envio, outro para recepção) é reservado exclusivamente para aquela conversa, garantindo uma taxa de bits constante e uma latência previsível, independentemente do tráfego na rede [Cunha et al. 2012]. O termo "multiplexação" refere-se justamente à capacidade de combinar múltiplos sinais digitais em um único meio de transmissão, intercalando amostras de cada sinal em seus respectivos tempos. O PABX IS 3000 é um exemplo de sistema que gerencia esses circuitos TDM em uma empresa. Ele se conecta à rede pública (PSTN) por meio de um enlace TDM padrão denominado **E1** (padrão europeu/brasileiro). Um enlace E1 opera a uma taxa de 2,048 Mbps e é estruturado em 32 *time slots* de 64 Kbps cada (canais DS0). Destes, 30 são usados para canais de voz (canais "B"), um é reservado para

sinalização (canal "D", geralmente o *slot* 16) e outro para sincronismo e alinhamento de quadro (*slot* 0) [Cunha et al. 2012]. A principal característica dessa arquitetura é a confiabilidade e a qualidade de serviço (QoS) garantidas pelo hardware, mas sua rigidez, a dependência de cabeamento dedicado e o alto custo de expansão a tornam uma infraestrutura legada, de difícil modernização e integração com aplicações de dados.

3.2. A Transição: Voz sobre IP (VoIP) e a Comutação de Pacotes

A Voz sobre IP (VoIP) representa uma mudança de paradigma fundamental, substituindo a comutação de circuitos pela *comutação de pacotes* (*packet switching*), a mesma arquitetura da Internet. No VoIP, o áudio analógico é digitalizado, comprimido por um *codec* (como G.711 ou Opus), e fragmentado em pequenos pacotes IP. Esses pacotes são enviados pela rede de dados sem um caminho dedicado; competem por largura de banda com outros tipos de tráfego e podem chegar fora de ordem, exigindo que o receptor os reordene e gerencie a variação no atraso (*jitter*). A comunicação VoIP é gerenciada por um conjunto de protocolos da camada de aplicação, sendo o SIP e o RTP os dois mais importantes definidos pela IETF (*Internet Engineering Task Force*).

Definido na RFC 3261, o SIP é o protocolo de controle e sinalização. Ele é um protocolo baseado em texto (similar ao HTTP/SMTP), responsável exclusivamente por "administrar" a sessão de comunicação. Sua função é localizar o usuário, iniciar a sessão (o "convite"), negociar os parâmetros da mídia (como quais *codecs* de áudio e vídeo serão usados, através do protocolo SDP - *Session Description Protocol*), modificar a sessão (ex: colocar em espera, transferir) e encerrá-la [Fernando 2025]. É importante notar que o SIP não transporta o áudio; ele apenas gerencia o estabelecimento da chamada.

Definido na RFC 3550, o RTP é um protocolo de mídia. Uma vez estabelecida a sessão pelo SIP, o RTP assume a responsabilidade de transportar, em tempo real, os dados de áudio (o *payload*). Ele opera geralmente sobre UDP (*User Datagram Protocol*) para minimizar a latência (não empregando a garantia de entrega do TCP, pois em voz, é preferível perder um pacote a retransmiti-lo com atraso). O cabeçalho RTP inclui informações críticas como *timestamps* (para sincronização e cálculo de *jitter*) e números de sequência (para detecção de perdas e reordenação de pacotes) [Chen et al. 2024].

3.3. Asterisk: O Gateway de Integração com a IA

O Asterisk é um *framework* de código aberto, criado pela Digium (agora Sangoma), que implementa um PABX completo em software. Sua arquitetura modular e flexível permite que ele interaja nativamente tanto com interfaces de telefonia TDM (por meio de placas de hardware específicas, como placas E1/T1) quanto com protocolos VoIP (SIP, IAX, H.323) [Sultan 2018]. Nesta arquitetura específica, o Asterisk não executa a lógica de IA localmente. Em vez disso, ele atua puramente como um *Media Gateway* e *Session Border Controller (SBC ou Controlador de Borda de Sessão)* simplificados. Sua função é converter a sinalização e a mídia no formato TDM (provenientes do PABX IS 3000 via E1/T1 ou de gateway físico) para o formato VoIP (SIP/RTP) compreensível pela rede de dados.

A conexão entre o Asterisk e a IA pode ser estabelecida por meio de um *SIP Trunk*. Diferentemente de uma API (Interface de Programas de Aplicação) REST (Transferência de Estados Representacional) tradicional, um SIP Trunk estabelece uma sessão de voz

bidirecional em tempo real. O Asterisk encaminha a chamada recebida do PABX IS 3000 diretamente para o endereço SIP do agente de IA (URI SIP), estabelecendo um canal de áudio RTP direto entre o chamador (no PABX) e o *bot* de voz (na nuvem) [Sultan 2018].

Para fins de validação e testes de latência, o usuário interage com o sistema por meio do *Softphone* (software emulador de telefone SIP), que se registra no Asterisk como um ramal IP padrão. Isso permite isolar a medição da latência introduzida pelo *gateway* e pela nuvem, removendo variáveis da rede telefônica pública.

3.4. Métricas de Desempenho: A Latência Conversacional

A validação de um sistema de voz também depende da sua usabilidade temporal. Em interações homem-máquina por voz, a métrica determinante da naturalidade é a **latência de troca de turno** (LTT ou *turn-taking latency*), que é o intervalo de silêncio entre o momento em que uma pessoa termina de falar e o momento em que a outra (ou um sistema) começa a responder. A psicolinguística e os estudos de interação humano-computador (HCI) estabelecem uma linha de base. Em conversas naturais em inglês, o tempo médio de resposta (o intervalo de silêncio entre o fim da fala de um interlocutor e o início da fala do outro) é de apenas **239 milissegundos** [Jacoby et al. 2024]. A distribuição estatística dessa latência define uma "janela de oportunidade" para uma resposta natural que varia de 280 ms (respostas antecipadas, ou *overlaps*) a **758 ms** [Jacoby et al. 2024]. Respostas que excedem esse limite superior de aproximadamente 0,7 segundos passam a ser percebidas como atrasos, gerando desconforto ou a percepção de falha na comunicação.

3.5. Desafio da Máquina e Limites de Tolerância

A arquitetura de IA, mesmo conectada via SIP direto, introduz latências intrínsecas (*buffer* de *jitter* do RTP, tempo de inferência do ASR/LLM/TTS). Pesquisas indicam que o tempo de processamento total da máquina, mesmo em sistemas otimizados, frequentemente situa-se entre **1000 ms e 1500 ms** (1 a 1,5 segundos), sem considerar os atrasos de rede (RTT) [Jacoby et al. 2024].

Embora a latência ideal seja inferior a 100 ms para a sensação de instantaneidade, diretrizes de HCI sugerem que atrasos de até **1000 ms** (1 segundo) em diálogos de máquina ainda mantêm o fluxo de pensamento do usuário, embora a "sensação de conversa fluida" seja perdida. Atrasos superiores a 2 segundos são considerados rupturas conversacionais graves [Jacoby et al. 2024].

Portanto, a análise de desempenho proposta neste trabalho medirá a **latência de ponta a ponta** — do fim da fala do usuário no *softphone* ao início do áudio de resposta — comparando-a com esses dois marcos: o ideal humano (< 758 ms) e o limite de tolerância da máquina (< 1500 ms), para determinar a viabilidade prática da arquitetura híbrida em um cenário real de atendimento.

4. Arquitetura Proposta

A arquitetura idealizada para a solução final busca integrar de forma digital a infraestrutura de telecomunicações legada com serviços de IA na nuvem, conforme ilustrado na Figura 1 e foi implementada na empresa VipTech Teleinformática LTDA.

A base instalada é composta pelo **PABX Digital TDM (IS 3000)**, equipamento central responsável por gerenciar ramais e chamadas. Sua tecnologia baseada em TDM

(*Time Division Multiplexing*) dificulta a conexão direta com redes IP modernas, mas sua preservação é crucial para a viabilidade econômica do projeto. A conexão com a rede pública (PSTN) é mantida por operadoras convencionais por meio de enlaces (*links*) digitais E1, garantindo qualidade na ”primeira milha”e recebendo chamadas de telefones fixos e celulares. A conexão com a rede de dados é feita por meio de canais IP SIP *trunk*, que utilizam o protocolo IP.

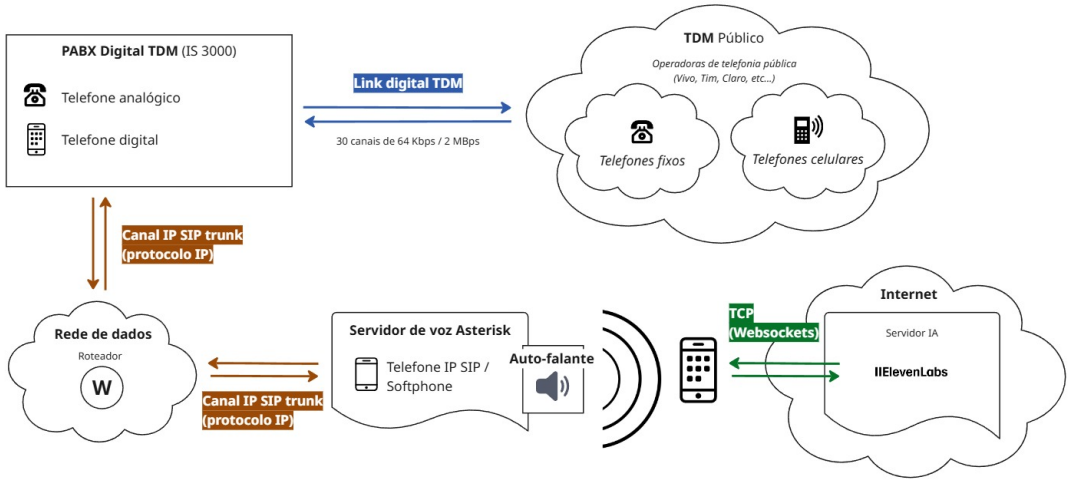

Figura 1: Cenário Experimental

O **Servidor Asterisk** atua como *gateway* de mídia, conectando-se ao PABX legado e participando do tráfego de voz e de sinalização via protocolo IP (SIP/RTP). A rede de dados deve ser dimensionada com mecanismos de QoS (*Quality of Service* ou Qualidade de Serviço) para priorizar pacotes de voz e minimizar a latência e o *jitter*. O foco é eliminar conversões analógicas desnecessárias. Na arquitetura ideal, a conexão entre o Asterisk e a IA é totalmente digital, por meio de **SIP Trunk**. Porém, devido a restrições de segurança de rede (firewall corporativo) na empresa onde o experimento foi realizado, a conexão SIP direta com a nuvem não foi testada.

A plataforma **ElevenLabs** atua como um orquestrador que permite acesso a diversos Grandes Modelos de Linguagem (LLMs) de mercado, como o ChatGPT (OpenAI) e o Gemini (Google), além de utilizar seus modelos proprietários. Neste fluxo, o Asterisk encaminha a chamada ao Servidor de IA, que gerencia o *pipeline* completo: transcreve a fala (ASR), processa a resposta com o LLM selecionado e sintetiza a voz (TTS) para retorno imediato [ElevenLabs 2025].

O segmento de telefonia inclui o PABX IS 3000, que manteve sua conexão direta com a rede pública e com o Asterisk via E1. O servidor Asterisk recebe as chamadas do PABX e as direciona para um *Softphone* na rede local. A interface de áudio adapta a saída de áudio do *Softphone*, conectando-a a um alto-falante de alta qualidade que reproduz a voz da rede telefônica. Um dispositivo móvel com o agente de IA capta esse áudio por meio de um microfone (”interface aérea”), processa a resposta e a reproduz. Essa configuração serviu como um “teste de estresse”, adicionando latência acústica e ruído ambiente. A premissa é que, se a IA funcionar nestas condições, o desempenho na conexão digital direta, que será objeto de futuros trabalhos, será superior.

Para ligar o mundo legado ao IP, foi criado um Tronco SIP entre o PABX e o Asterisk. No PABX IS 3000 foi criada uma rota de saída apontando para o IP do Asterisk e ajustado o plano de numeração. O servidor Asterisk foi configurado com endereço IP estático. No arquivo `pjsip.conf`, o PABX foi definido como *endpoint* (`type=friend`), forçando o uso de codecs G.711 (alaw/ulaw) [International Telecommunication Union 1988] para manter a qualidade de áudio sem compressão adicional. Na configuração do Tronco SIP com o ElevenLabs, definiram-se os parâmetros de autenticação. O número de telefone e a identificação com DDI e DDD. Para as credenciais SIP, utilizam-se o nome de usuário e a senha. O *endpoint* de saída foi o endereço de destino para roteamento do SIP INVITE. Devido às restrições de rede, o *softphone* (MicroSIP [MicroSIP nd]) atuou como ponto de terminação. Assim, o servidor SIP foi mapeado para o endereço IP do Asterisk. O usuário empregou o ramal 2000. O protocolo da camada de transporte foi o UDP, de modo a minimizar a latência, conforme o RFC 3551 [Schulzrinne and Casner 2003].

5. *Pipeline* Proposto de IA Conversacional (via SIP)

Diferentemente das integrações baseadas em AGI local, nesta arquitetura o processamento de IA ocorre inteiramente em uma plataforma externa, conectada via SIP. O Asterisk entrega o fluxo de áudio RTP bruto ao provedor de IA, que executa o *pipeline* internamente. Este modelo, conhecido como *Voice AI over SIP*, consolida as etapas de processamento para reduzir a latência de transporte.

- **Recepção de Áudio (SIP/RTP):** O agente de IA atende à chamada SIP como se fosse um telefone comum e passa a receber o fluxo de áudio RTP do usuário.
- **ASR (*Speech-to-Text*) e VAD (*Voice Activity Detection* Detecção de Atividades de Voz):** A plataforma de IA processa o *stream* RTP de entrada. Algoritmos de VAD detectam quando o usuário começa e para de falar. O áudio é transcrito em tempo real (*streaming recognition*). A eficiência do VAD é crítica aqui: ele precisa distinguir o ruído de fundo da fala e decidir o momento exato de "cortar" a escuta para enviar o texto ao modelo de linguagem [ElevenLabs 2025].
- **LLM (*Large Language Model*) como Cérebro:** O texto transcrito é enviado diretamente para um Grande Modelo de Linguagem (como GPT-4o ou modelos LLaMA otimizados). Diferentemente dos sistemas de NLU tradicionais baseados em intenções rígidas (como o Dialogflow), o LLM recebe o texto conversacional e gera uma resposta completa em linguagem natural, mantendo o contexto da conversa. Esta abordagem permite diálogos mais fluidos e menos roteirizados [Chen et al. 2024].
- **TTS (*Text-to-Speech*) com *Streaming*:** A resposta gerada pelo LLM é convertida em áudio. Para minimizar a latência, a síntese de voz utiliza *streaming*: o áudio começa a ser enviado de volta via RTP para o Asterisk assim que os primeiros *tokens* de texto são gerados pelo LLM, sem esperar a frase completa. Isso permite que o usuário comece a ouvir a resposta quase imediatamente [Chowdhury and Hussan 2023, ElevenLabs 2025].

6. O Experimento

O objetivo do experimento é validar se a plataforma consegue manter uma conversa fluida com humanos. Todas as frases de teste são perguntas, o que obriga a IA a processar a fala

(ASR), entender o contexto (NLU) e gerar *output*. Como métricas, foram usadas a Taxa de Erro de Palavras (WER, do inglês *Word Error Rate*), que avalia a precisão da transcrição e é definida como a razão entre o número de erros (trocas, inserções e cortes) e o total de palavras, e a Latência de Troca de Turno (Delta T), que é o tempo de silêncio entre a pergunta do usuário e a resposta da IA. Latências acima de 758 ms passam a ser percebidas como atrasos, e as acima de 1500 ms prejudicam a naturalidade [Jacoby et al. 2024]. A análise focou na média e no "pior caso", utilizando Percentil 95, métrica que considera apenas os 95% mais baixos de um conjunto de dados, descartando picos.

Para garantir a robustez dos resultados, definiu-se um *Corpus Experimental* composto por 20 frases em português, divididas em duas categorias. Cada frase foi executada 2 vezes por 2 indivíduos distintos, totalizando 80 amostras.

O grupo de frases simples (baixa ambiguidade) avalia o desempenho básico do sistema em condições ideais e é descrito na Tabela 1. Já o grupo de frases complexas e adversas introduz desafios, como ambiguidade fonética, vocabulário raro e ruído branco (C15 e C16), para testar os limites do ASR, conforme descrito na Tabela 2.

Tabela 1: Listagem de perguntas simples

ID	**Tipo de Pergunta**	**Texto de Referência (GT)**
S01	Afirmação c/ Interrogação	O céu é azul e o mar é verde escuro?
S02	Pergunta Simples	Qual a capital do Brasil atualmente?
S03	Comando de Listagem	Você pode listar os três maiores rios da Europa?
S04	Sequência de Números	Meu código é um, três, nove, zero, sete?
S05	Negação	Isso não faz sentido na minha opinião, certo?
S06	Dúvida Formal	Eu poderia receber uma explicação mais detalhada sobre isso?
S07	Composição Simples	Eu gosto de café, chá gelado e água, e você?
S08	Expressão Temporal	O relatório deve ser entregue na próxima terça-feira?
S09	Foco em Pronomes	Eles e nós vamos participar da reunião final?
S10	Vocabulário Neutro	A temperatura hoje está agradável na cidade?

7. Resultados e Discussão

A precisão do sistema na transcrição da fala do usuário é o primeiro indicador de viabilidade, pois falhas nesta etapa comprometem todo o fluxo conversacional. A métrica utilizada foi a Taxa de Erro de Palavras (WER, do inglês *Word Error Rate*). Os testes foram realizados com um *corpus* de 20 frases, divididas em "Simples" (S01-S10) e "Complexas/Adversas" (C11-C20), totalizando 80 amostras processadas (4 rodadas de teste por frase). A Tabela 3 apresenta os resultados obtidos, em que o conjunto S01-S10 apresenta 0,5% de erro, enquanto o conjunto C11-C20 apresenta 2,2%.

A latência é o fator crítico para a naturalidade da conversa. Verificou-se se o sistema conseguiria operar dentro dos limites aceitáveis para interação humana ($<$1500ms) e quão próximo chegaria do ideal humano ($<$758ms). A Tabela 4 apresenta os resultados da Latência de Troca de Turno (LTT), medida em milissegundos (ms), para todas as 80 interações.

Tabela 2: Listagem de perguntas complexas

ID	Condição	Texto de Referência (GT)
C11	Normal	Eu e a Gabriela visitamos a Biblioteca Nacional, certo?
C12	Normal	O mandato dele foi revogado por um mal entendido, é isso?
C13	Normal	A sinestesia é um conceito filosófico complexo?
C14	Normal	Se chover, devo levar o guarda-chuva, ou usar o casaco leve?
C15	Ruído Forte	A logística para o evento está confirmada, certo?
C16	Ruído Forte	Precisamos checar o feedback da equipe de marketing imediatamente?
C17	Normal	A placa do carro é B X I sete três zero zero?
C18	Normal	Embora ele tenha tentado se justificar, a decisão final já estava tomada, não é?
C19	Normal	Eu quero, eu quero o valor total, porém com o desconto?
C20	Normal	”Onde há fumaça, há fogo”, disse o detetive, o que ele quis dizer?

Tabela 3: Resumo dos Resultados de WER por Categoria

Categoria	Palavras	Erros	WER (%)	Observações
Simples (S01-S10)	83	1	1,20%	Corte da palavra “chá” da frase S07
Complexas (C11-C20)	105	4	3,80%	1 substituição e 3 adições. De “guarda-chuva” para “água da chuva” na frase C14
Total Geral	**188**	**5**	**2,65%**	**Desempenho global de alta precisão.**

A análise inicial dos dados brutos (Média: 1.440 ms) sugere que a latência total do sistema excede o limiar ideal de 758 ms para uma conversa humana natural e situa-se ligeiramente acima da zona de conforto de 1500 ms para agentes virtuais [Jacoby et al. 2024]. No entanto, uma análise mais aprofundada da composição dessa latência revela que o gargalo não está na Inteligência Artificial, mas na infraestrutura legado de teste.

Para compreender a origem do atraso, foi realizada a decomposição dos tempos médios gastos em cada etapa do *pipeline*. Os dados, apresentados na Tabela 5, separam o tempo de processamento da IA do tempo de transporte na rede telefônica.

O resultado mais crítico é o tempo médio de **600 ms** gasto na primeira etapa, que representa quase metade (41,7%) de toda a latência do sistema e é composto por fatores alheios à inteligência artificial: propagação na rede TDM, comutação no PABX e no Asterisk, buffers de *jitter* e, principalmente, a latência acústica da configuração “interface aérea”. Este atraso atua como um “piso” intransponível para o experimento.

Tabela 4: Estatística Descritiva da Latência (ms)

Métrica	**Valor (ms)**	**Ref. Humana**	**Status**
Média	1.440 ms	$\sim$239 ms	Acima do ideal humano
Mediana	1.390 ms	-	-
Mínimo	1.203 ms	-	Próximo do limite de tolerância
Máximo	2.127 ms	-	Perceptivelmente lento
Desvio Padrão	174 ms	-	Baixa variabilidade (Estável)
Percentil 95 (P95)	1.809 ms	$<$758 ms	Aceitável para máquinas ($<$2000ms)

Tabela 5: Decomposição Média da Latência por Etapa

Etapa do Pipeline	**Tempo (ms)**	**%**	**Natureza**
1. Rede (TDM público $\rightarrow$ Softphone)	600 ms	41,7%	Infraestrutura (Fixo)
2. Reconhecimento de Fala (ASR)	139 ms	9,7%	Processamento IA
3. Processamento Cognitivo (LLM)	484 ms	33,7%	Processamento IA
4. Síntese de Voz (TTS)	215 ms	15,0%	Processamento IA
Latência Total (LTT)	**1.438 ms**	**100%**	

Ao isolar o tempo de processamento da IA (ASR + LLM + TTS), obtém-se um tempo médio de **838 ms**. Este valor é bem competitivo e está muito próximo da janela ideal de conversação humana ($<$758 ms), o que demonstra que o gargalo da solução não é o processamento cognitivo.

Os resultados obtidos foram confrontados com as normas regulatórias de telecomunicações e padrões da indústria. O resultado de **2,65% de WER** situa-se em uma faixa de excelência [McGillivray 2024]. Este desempenho deve-se à arquitetura dos codecs adotados. A configuração do Asterisk utiliza o codec **G.711** [International Telecommunication Union 1988], que oferece uma amostragem sem perda de compressão na banda de voz. Ao evitar o uso de codecs de alta compressão (como G.729), a solução garantiu que o motor de IA recebesse um áudio acústico íntegro.

A latência média total de 1.440 ms excede o limiar de 239 ms para conversação humana [Jacoby et al. 2024]. Contudo, a análise isola a infraestrutura de rede como fator de violação. O componente de rede (600 ms) ultrapassa, isoladamente, o limite de 400 ms para o atraso unidirecional da recomendação ITU-T G.114 [International Telecommunication Union 2003]. O valor também excede as metas da Anatel para redes terrestres ($\sim$80 ms) [Agência Nacional de Telecomunicações 2011]. Isso confirma que a degradação é um artefato do ambiente de teste, e não uma limitação da IA. Porém, ao projetar a migração para um ambiente de produção com SIP Trunking direto (removendo a interface aérea e utilizando WebSockets [ElevenLabs 2025]), o orçamento de latência é reestruturado. Nesse caso, estima-se uma redução de 600 ms para $\sim$100 ms (considerando a pacotização RTP de 20 ms, conforme o RFC 3551 [Schulzrinne and Casner 2003]). Sendo a eliminação da "interface aérea"(microfone/alto-falante) e dos *buffers* de *jitter* intermediários do softphone os principais responsáveis por essa economia de tempo. Já o processamento IA é mantido em 838 ms. E o total projetado é de $\sim$938 ms. Este valor coloca a solução abaixo da barreira de 1 segundo, na zona de "tolerância cognitiva" aceitável.

A análise segregada permite concluir que a arquitetura híbrida é tecnicamente viável. O *pipeline* de IA mostrou-se rápido (838 ms) e preciso (WER=2,65%) [McGillivray 2024]. O tempo de "pensamento" do LLM (484 ms) é o componente variável mais longo, mas aceitável. Conclui-se que a modernização do PABX legado é viável, mas a melhoria da camada de rede (migração para SIP Trunk puro na última milha) é mandatória para desbloquear todo o potencial de velocidade da Inteligência Artificial.

8. Conclusão e Trabalhos Futuros

O presente trabalho dedicou-se a investigar a viabilidade técnica e o desempenho de uma plataforma híbrida de atendimento ao cliente, projetada para integrar infraestruturas de telefonia legada (PABX TDM) com serviços modernos de Inteligência Artificial Conversacional baseados em nuvem. A motivação central residiu no desafio de modernizar sistemas de voz tradicionais sem a necessidade de substituição total do hardware existente, uma demanda recorrente no mercado corporativo.

A arquitetura proposta, que utiliza o servidor Asterisk como *gateway* de orquestração e a plataforma ElevenLabs como motor de IA via *SIP Trunking*, demonstrou ser funcional e robusta. A decomposição dos tempos de processamento mostra que o principal fator de atraso esteve associado à infraestrutura de teste e não ao pipeline de IA. Isso indica que, em um cenário de implantação totalmente digital, a solução tende a operar dentro de limites aceitáveis de usabilidade na interação homem-máquina. Conclui-se, portanto, que a arquitetura híbrida proposta é tecnicamente viável e representa uma estratégia realista para a modernização gradual de sistemas legados de voz, combinando a preservação do investimento existente com a adoção de tecnologias baseadas na nuvem.

Como trabalhos futuros, pretende-se validar a arquitetura em ambiente totalmente digital com SIP Trunk direto, realizar testes de escalabilidade com múltiplas chamadas simultâneas, investigar o impacto de diferentes codecs na precisão do reconhecimento de fala, implementar suporte à interrupção de fala para interação full-duplex e conduzir análise econômica comparativa entre a solução proposta e modelos tradicionais de atendimento, visando consolidar sua viabilidade técnica e operacional em cenários reais.

Agradecimentos

Este trabalho foi parcialmente apoiado pela Loomi, Porto Digital, Recife-PE, Brasil.

Referências

Agência Nacional de Telecomunicações (2011). Resolução nº 574, de 28 de outubro de 2011. aprova o regulamento de gestão da qualidade do serviço de comunicação multimídia (rgq-scm). ANATEL.

Chen, J., Wei, Y., Lin, Z., Tan, X., Wang, B., and Zhang, L. (2024). Wavchat: A survey of spoken dialogue models. arXiv:2411.13577. Disponível em: `https://arxiv.org/abs/2411.13577`.

Chowdhury, M. Q. Z. and Hussan, M. A. (2023). A review-based study on different text-to-speech technologies. *arXiv:2312.11563*. Disponível em: `https://www.semanticscholar.org/paper/A-review-based-study-on-different-Text-to-Speech-Chowdhury-Hussan/49b56a07fa9193811cc411f09020a3723fb515ed`.

Cunha, M. A. P., Silva, K. S., Mota, D. F. M., and Vasconcellos, A. A. (2012). Uma arquitetura modular de hardware e software para pabx voip baseado em asterisk. In *XXX Simpósio Brasileiro de Telecomunicações (SBT)*, Brasília, DF.

ElevenLabs (2025). Sip trunking integration for agents. Disponível em: https://elevenlabs.io/agents/integrations/sip-trunking.

Fernando, R. (2025). Voice driven ai based automated voip pbx for smes. In *International Conference on Advanced Computing Technologies (ICACT 2025)*, Sri Lanka.

González-Bustamante, B. (2024). Textclass benchmark: A continuous elo rating of llms in social sciences. arXiv:2412.00539. Disponível em: https://arxiv.org/abs/2412.00539.

Hugging Face (2025). Tts arena and leaderboards. Disponível em: https://tts-agi-tts-arena-v2.hf.space/leaderboard.

International Telecommunication Union (1988). Recommendation g.711: Pulse code modulation (pcm) of voice frequencies. Recommendation, ITU-T, Geneva.

International Telecommunication Union (2003). Recommendation g.114: One-way transmission time. Recommendation, ITU-T, Geneva.

Jacoby, D., Zhang, T., Mohan, A., and Coady, Y. (2024). Human latency conversational turns for spoken avatar systems. arXiv:2404.16053. Disponível em: https://arxiv.org/abs/2404.16053.

Madsen, L., Van Meggelen, J., and Bryant, R. (2013). *Asterisk: The Definitive Guide*. O'Reilly Media, 4th edition.

McGillivray, B. (2024). Speech-to-text api benchmarks: Accuracy, speed, and cost compared. Disponível em: https://deepgram.com/learn/speech-to-text-benchmarks.

MicroSIP (n.d.). Microsip help and documentation. Disponível em: https://www.microsip.org/help.

Minixhofer, C., Klejch, O., and Bell, P. (2024). Ttsds – text-to-speech distribution score. arXiv:2407.12707. Disponível em: https://arxiv.org/abs/2407.12707.

Schulzrinne, H. and Casner, S. (2003). Rtp profile for audio and video conferences with minimal control. RFC 3551, Internet Engineering Task Force.

Sultan, P. (2018). Asterisk eagi with google speech recognition. Disponível em: https://github.com/phsultan/asterisk-eagi-google-speech-recognition.

Unify Software and Solutions (2020). *Is3000 / SIP@Net: BIM Manual*. Disponível via Scribd. Disponível em: https://pt.scribd.com/document/464265931/BIM-Manual1579853077.

VoiceWriter (2025). Automatic speech recognition leaderboards. Disponível em: https://voicewriter.io/speech-recognition-leaderboard.

VoIP Info (2025). Voip-info.org - what is voip. Disponível em: https://www.voip-info.org/what-is-voip/.