O codec pode dobrar o erro do seu bot de voz

Estudo de degradação por codec de telefonia no reconhecimento de fala em português — 9 condições de áudio, 3 arquiteturas de modelo, dados medidos

Julho de 2026

Resumo executivo

Todo voicebot, IVR com IA ou análise de chamadas escuta o cliente **através de um codec de telefonia** — e a escolha desse codec costuma ser feita pela equipe de rede olhando banda e custo, sem ninguém medir o efeito sobre o reconhecimento de fala. Este estudo mede exatamente isso.

Transcodificamos um corpus de fala espontânea em português brasileiro (CORAA NURC-SP, 500 trechos com referência humana) através da cadeia real de codecs de telefonia — G.711, G.722, G.729, AMR-NB, AMR-WB, Opus e MP3 8 kHz — e medimos o WER de três arquiteturas de reconhecimento em cada condição, incluindo o **pulse-precision**, o motor da SipPulse AI.

ACHADO	NÚMERO
G.729 quase dobra o erro — em todos os modelos testados	WER 15.5% → 28.9% (1.9×) no pulse-precision; padrão idêntico em Whisper e Parakeet
AMR-NB (voz 3G) é o segundo pior	+6.5 p.p. de WER em média
Codecs wideband são quase de graça	G.722 +0.7 · Opus +1.0 p.p. — indistinguíveis do áudio limpo
O clássico G.711 custa ~3 pontos	+2.8 p.p. (A-law) — o preço do narrowband 8 kHz
Benchmarks de fala lida escondem o problema	no MLS (leitura), o G.729 custa +2.2 p.p.; na fala espontânea, +12.3 p.p. — mais de 5× o dano

A conclusão prática: antes de trocar de fornecedor de STT, olhe o codec do trecho SIP que alimenta o bot. **Migrar o leg de G.729 para G.711, G.722 ou Opus recupera mais precisão do que qualquer troca de modelo.**

1. Por que medir isso

O áudio que chega a um motor de reconhecimento de voz numa operação de telefonia nunca é o áudio que saiu da boca do cliente. Ele atravessou pelo menos um codec — escolhido pela operadora, pelo tronco SIP ou pela configuração do PBX — que comprime a voz para caber em banda de telefonia. Cada codec joga fora uma parte diferente do sinal.

Quem opera voicebots conhece o sintoma: o mesmo bot que vai bem num cliente vai mal no outro, e a diferença não está no modelo — está no caminho do áudio. Mas esse efeito quase nunca é quantificado, porque exigiria o **mesmo áudio, com a mesma referência humana, atravessando cada codec** — algo que não existe em gravações reais, que já chegam degradadas por um caminho desconhecido.

Nós construímos essa medição.

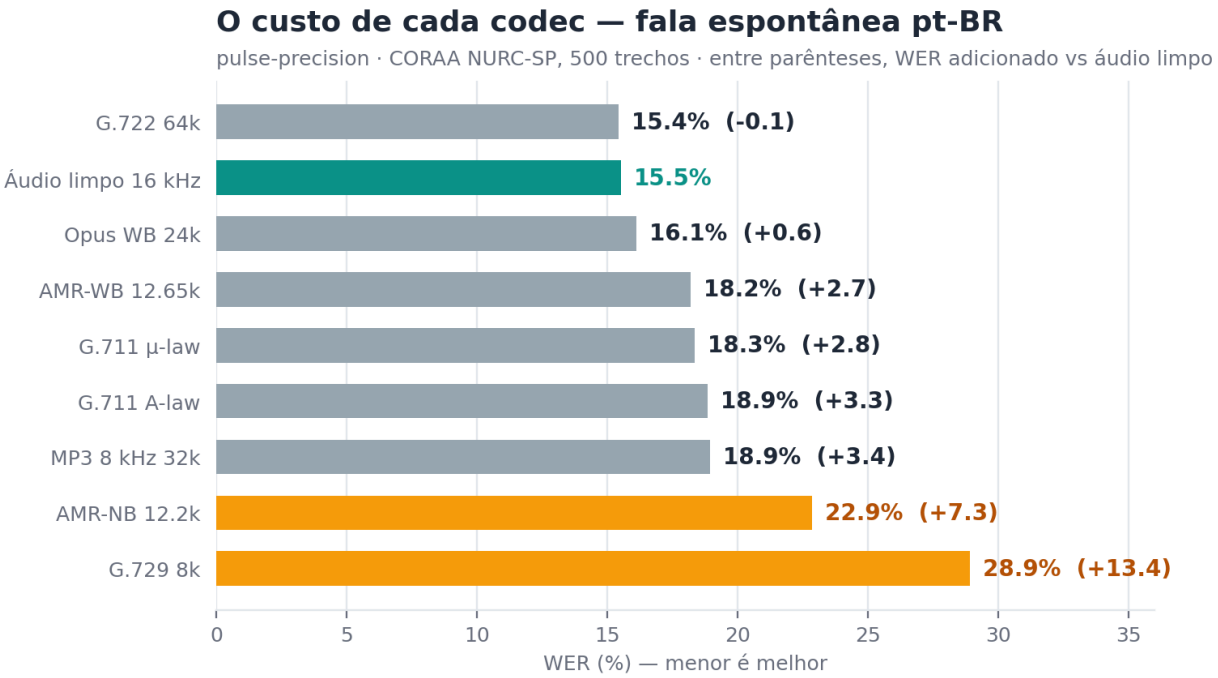
2. Metodologia

- **Fonte controlada:** CORAA NURC-SP, o corpus público de fala espontânea em português brasileiro com transcrição verificada por humanos — 500 trechos (48 min de áudio), nativos em ≥16 kHz.
- **Degradação real, não simulada por filtro:** cada trecho foi **transcodificado de verdade** por cada codec na sua taxa nativa (G.711 A-law/μ-law 64k, G.722 64k, AMR-NB 12.2k, AMR-WB 12.65k, Opus WB 24k VoIP, MP3 8 kHz 32k) e decodificado de volta para 16 kHz — exatamente a perda que uma perna de telefonia impõe. O G.729 (8k Annex B) passou por um encoder/decoder de referência dedicado, já que não há suporte nativo nas ferramentas usuais.
- **Controle de fala lida:** o mesmo protocolo aplicado ao MLS português (fala lida de audiolivros) para separar o efeito do codec da dificuldade da fala espontânea.
- **Métrica:** WER com normalização pt-BR (caixa, pontuação, números) aplicada igualmente à referência e a todas as hipóteses. Mesma cadeia de áudio para todos os modelos.
- **Modelos:** pulse-precision (SipPulse AI), OpenAI Whisper large-v3 e NVIDIA Parakeet-TDT 0.6B v3 — três arquiteturas distintas, para separar o efeito do codec de idiossincrasias de um modelo só.

> **O que este protocolo não modela:** perda de pacote e PLC, cross-talk entre falantes, música de espera e acústica de viva-voz. Ele isola **apenas** o dano do codec — o número real em produção soma esses outros fatores.

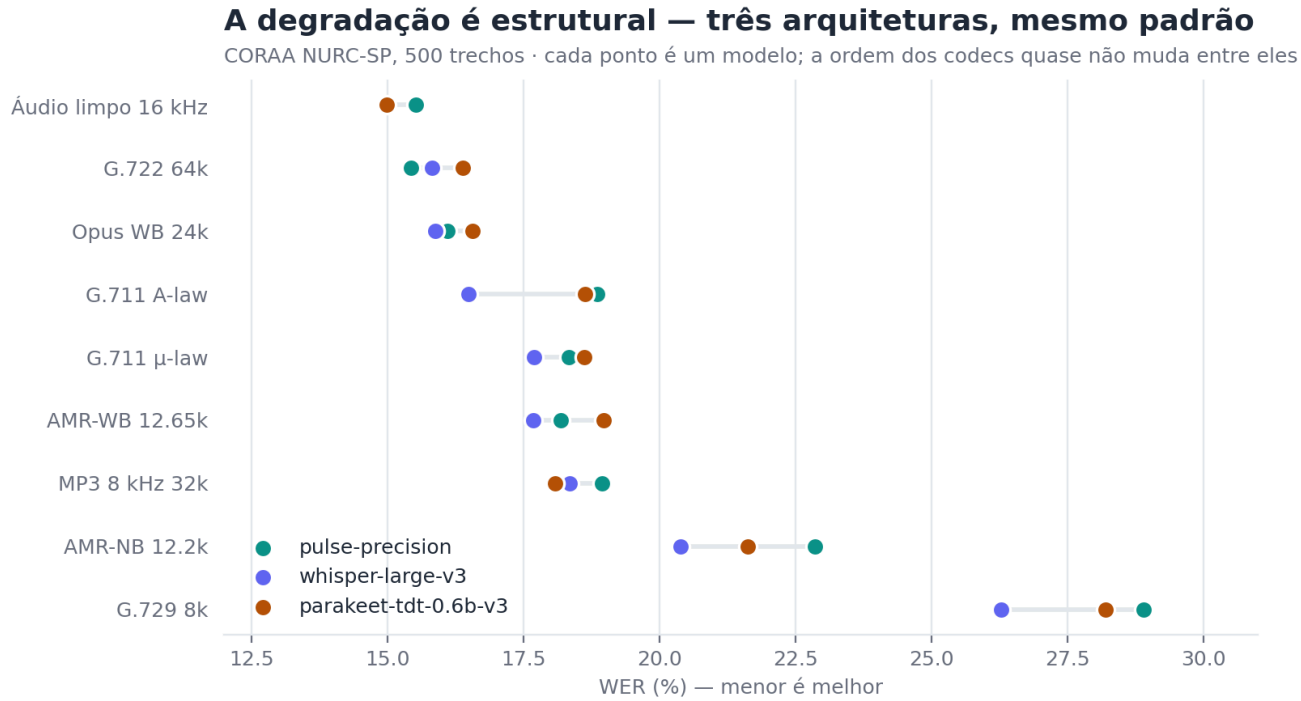
3. Resultados

O custo de cada codec



A leitura é direta: **G.722 e Opus empatam com o áudio limpo** — a banda larga preserva o que o modelo precisa. Os narrowband clássicos (G.711, MP3 8k) custam ~3 pontos. O **AMR-NB** custa 6 pontos. E o **G.729** — o codec favorito de quem economiza banda — **quase dobra o WER**.

Três arquiteturas, o mesmo padrão

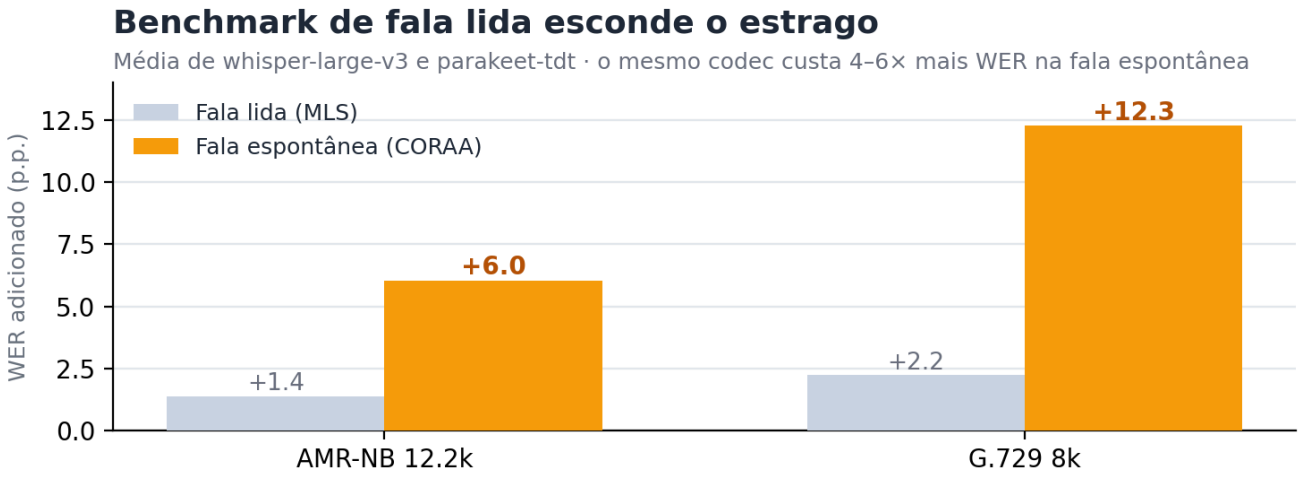


MODELO	LIMPO	G.711 A-LAW	G.711 M-LAW	MP3 8 KHZ 32K	G.722 64K	AMR-NB 12.2K	AMR-WB 12.65K	OPUS WB 24K	G.729 8K
pulse-precision	15.5%	18.9% (+3.3)	18.3% (+2.8)	18.9% (+3.4)	15.4% (-0.1)	22.9% (+7.3)	18.2% (+2.7)	16.1% (+0.6)	28.9% (+13.4)
pulse-precision-robusto	17.0%	18.3% (+1.3)	18.8% (+1.9)	18.9% (+1.9)	16.0% (-0.9)	22.6% (+5.6)	18.6% (+1.6)	17.4% (+0.4)	27.8% (+10.9)
whisper-large-v3	15.0%	16.5% (+1.5)	17.7% (+2.7)	18.4% (+3.4)	15.8% (+0.9)	20.4% (+5.4)	17.7% (+2.7)	15.9% (+0.9)	26.3% (+11.3)
parakeet-tdt-0.6b-v3	15.0%	18.6% (+3.6)	18.6% (+3.6)	18.1% (+3.1)	16.4% (+1.4)	21.6% (+6.6)	19.0% (+4.0)	16.6% (+1.6)	28.2% (+13.2)

Entre parênteses, pontos de WER adicionados sobre o áudio limpo do mesmo modelo. pulse-precision-robusto = perfil de decodificação do pulse-precision otimizado para canais degradados.

O padrão se repete nas três arquiteturas — a degradação é **propriedade do codec, não do modelo**. Nenhum fornecedor escapa: se o áudio chega em G.729, qualquer motor do mercado erra quase o dobro. O perfil **robusto** do pulse-precision recorta o pior caso (−1,1 p.p. no G.729, −0,6 no G.711) ao custo de ~1,5 p.p. no áudio limpo — uma troca que faz sentido quando o tráfego é majoritariamente narrowband.

4. Por que ninguém fala disso: fala lida esconde o estrago



Nos benchmarks públicos — quase todos de fala lida ou preparada — o efeito do codec parece pequeno: +1 a +2 pontos no pior caso. Na fala espontânea de uma ligação real, o mesmo codec custa **4 a 6× mais**. A explicação: a fala lida é redundante e previsível, então o modelo reconstrói o que o codec destruiu; na fala espontânea — hesitações, sobreposições, prosódia de conversa — não há redundância que salve.

MODELO	LIMPO	G.711 A-LAW	G.711 M-LAW	MP3 8 KHZ 32K	G.722 64K	AMR-NB 12.2K	AMR-WB 12.65K	OPUS WB 24K	G.729 8K
mls/whisper-large-v3	6.3%	7.0% (+0.7)	7.0% (+0.7)	6.5% (+0.2)	6.3% (+0.1)	7.5% (+1.2)	6.1% (-0.2)	6.3% (-0.0)	7.7% (+1.4)
mls/parakeet-tdt-0.6b-v3	6.0%	6.7% (+0.6)	6.6% (+0.6)	6.7% (+0.6)	6.1% (+0.0)	7.6% (+1.6)	8.2% (+2.1)	6.5% (+0.5)	9.1% (+3.1)

MLS português (fala lida), mesmas condições. O WER absoluto é baixo e a degradação, leve — o cenário que os rankings públicos mostram.

5. Recomendações práticas

CODEC	WER ADICIONADO (MÉDIA DE 3 MODELOS)	SEVERIDADE
G.722 64k	+0.7 p.p.	**≈ de graça**
Opus WB 24k	+1.0 p.p.	**≈ de graça**
G.711 A-law	+2.8 p.p.	**moderado**
G.711 μ-law	+3.1 p.p.	**moderado**
AMR-WB 12.65k	+3.1 p.p.	**moderado**
MP3 8 kHz 32k	+3.3 p.p.	**moderado**
AMR-NB 12.2k	+6.5 p.p.	**alto**
G.729 8k	+12.6 p.p.	**crítico**

- **Negocie o codec do leg do bot.** O trecho SIP que alimenta o STT pode usar codec diferente do resto da chamada. G.711, G.722 ou Opus nesse leg recuperam quase toda a precisão — **sem custo de modelo**.
- **Se o tronco entrega G.729, evite re-transcodificar** para "melhorar": o dano já foi feito na primeira codificação; transcodificações extras só somam perda.
- **Dimensione a expectativa por codec.** Um bot medido em laboratório com áudio limpo vai errar ~2 p.p. a mais em G.711 e ~13 p.p. a mais em G.729. Isso muda metas de contenção e SLAs.
- **Para tráfego majoritariamente narrowband,** o pulse-precision oferece o perfil **robusto**, que reduz o dano nos codecs agressivos.
- **Vá além do codec quando possível:** o modelo pode ser ajustado com áudio do seu próprio tronco (fine-tuning por domínio), absorvendo o canal específico da sua operação.

6. Conclusão

A precisão de um sistema de voz não é só do modelo — **é do caminho do áudio**. Este estudo mostra, com o mesmo áudio e a mesma referência humana, que a escolha do codec move o WER mais do que a escolha entre os principais motores do mercado: de **quase nada** (G.722, Opus) a **quase o dobro do erro** (G.729). Quem trata o codec como detalhe de rede está deixando precisão — e contenção — na mesa.

> **Quer saber quanto o seu tronco está custando em precisão?** Mande 30 minutos de chamadas do seu tráfego real e devolvemos o diagnóstico: WER estimado por codec e o ganho esperado ao mudar o leg do bot. **contato@sippulse.ai · sippulse.ai**

Metodologia e resultados completos disponíveis sob solicitação. SipPulse AI · julho de 2026.

Apêndice — tabela completa

MODELO	LIMPO	G.711 A-LAW	G.711 M-LAW	MP3 8 KHZ 32K	G.722 64K	AMR-NB 12.2K	AMR-WB 12.65K	OPUS WB 24K	G.729 8K
pulse-precision	15.5%	18.9% (+3.3)	18.3% (+2.8)	18.9% (+3.4)	15.4% (-0.1)	22.9% (+7.3)	18.2% (+2.7)	16.1% (+0.6)	28.9% (+13.4)
pulse-precision-robusto	17.0%	18.3% (+1.3)	18.8% (+1.9)	18.9% (+1.9)	16.0% (-0.9)	22.6% (+5.6)	18.6% (+1.6)	17.4% (+0.4)	27.8% (+10.9)
whisper-large-v3	15.0%	16.5% (+1.5)	17.7% (+2.7)	18.4% (+3.4)	15.8% (+0.9)	20.4% (+5.4)	17.7% (+2.7)	15.9% (+0.9)	26.3% (+11.3)
parakeet-tdt-0.6b-v3	15.0%	18.6% (+3.6)	18.6% (+3.6)	18.1% (+3.1)	16.4% (+1.4)	21.6% (+6.6)	19.0% (+4.0)	16.6% (+1.6)	28.2% (+13.2)
faster-whisper-large-v3	15.5%	18.6% (+3.1)	18.7% (+3.2)	19.4% (+3.9)	15.6% (+0.1)	23.4% (+7.9)	18.6% (+3.1)	16.5% (+1.0)	29.0% (+13.5)
whisper-sem-vad	29.9%	19.5% (-10.3)	27.9% (-2.0)	44.9% (+15.0)	21.5% (-8.4)	35.5% (+5.6)	27.1% (-2.7)	25.2% (-4.7)	63.4% (+33.5)

faster-whisper-large-v3 = mesma rede do Whisper large-v3 com runtime otimizado (CTranslate2). whisper-sem-vad = Whisper large-v3 decodificando o áudio bruto, sem detecção de voz prévia — incluído para mostrar que a instabilidade do pipeline (alucinação em silêncio/ruído) domina o efeito do codec: o WER limpo já parte de 30% e chega a 63% no G.729. Números de fala lida (MLS) na seção 4.