WIGVO: Real-Time Bidirectional Speech Translation over Legacy PSTN Calls via Dual-Session Echo Gating

Hyeong-seob Kim **Sang-Woo Son** **Hyun-woo Cho** **Hyeonsang Kim** **Jinmo Kim**

WIGTN, Seoul, Republic of Korea
harrison@wigtn.com

Abstract

Real-time speech translation with large language models (LLMs) has become feasible in controlled wideband settings—mobile apps, web browsers, and end-to-end full-duplex systems pushing latency below 200 ms—where developers can assume client-side echo cancellation. However, deploying such systems over the Public Switched Telephone Network (PSTN) remains challenging due to narrowband G.711 audio, unpredictable round-trip delays, and absence of client-side signal processing. We present **WIGVO** (WIGTN Voice-Only), a server-side relay system that enables bidirectional LLM-based speech translation over ordinary telephone calls without requiring app installation or carrier integration. A central contribution is addressing what we term *echo-induced self-reinforcing translation loops*: synthesized speech echoing back through the PSTN gets re-ingested and repeatedly translated. WIGVO solves this through a dual-session architecture with deterministic silence injection and energy-based voice activity detection (VAD) gating. We evaluate WIGVO on 155 Korean–English PSTN calls (148 instrumented, 147 completed) across three communication modes—voice-to-voice (V2V), text-to-voice (T2V), and full-agent—observing 555 ms median caller-to-callee latency and 2,684 ms median callee-to-caller latency, zero echo-induced translation loops, COMET semantic adequacy of 0.71 (en→ko) and 0.62 (ko→en) against offline LLM references, and USD 0.28 per minute cost. The system is deployed at `https://wigvo-web-gzjzn35jyq-du.a.run.app`, with a video walkthrough at `https://youtu.be/_ixVEnHJxjk`. Evaluation scripts and anonymized call logs are available in the open-source repository.

1 Introduction

Streaming speech-to-speech translation has advanced rapidly (Seamless Communication et al., 2023a; OpenAI, 2024), but existing systems assume wideband audio and client-side acoustic echo cancellation (AEC) (ITU-T, 2015). The Public Switched Telephone Network (PSTN) remains the primary inbound interface for local businesses, hospitals, and government offices (International Telecommunication Union, 2020), yet it encodes audio via G.711 μ-law at 8 kHz (ITU-T, 1988), introduces 80–600 ms delays (ITU-T, 2003), and provides no AEC. For foreign residents, people with speech anxiety, or users with hearing impairments, making such calls remains a significant barrier.

We present **WIGVO**, a server-side relay that bridges a web client and a standard PSTN phone number through two concurrent LLM-backed streaming sessions. The caller speaks (or types) via a browser; the callee answers on an ordinary phone—no app installation, carrier integration, or special hardware is required on either end. WIGVO translates bidirectionally in real time, supporting voice-to-voice (V2V), text-to-voice (T2V), and full-agent modes.

A key challenge is what we call an **echo-induced self-reinforcing translation loop**: synthesized speech echoing back through the telephone network gets re-recognized and repeatedly translated, creating a runaway feedback cycle (Section 3.1). Our solution combines three architectural mechanisms—directional separation, deterministic silence injection, and energy-based gating—described in Section 3.2.

We evaluate WIGVO on 147 completed PSTN calls and report latency, echo suppression, and cost metrics. Our contributions are:

- We formalize the *echo-induced translation loop* problem in streaming speech translation over telephony.
- We propose a *dual-session gated architecture* combining directional separation, silence in-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 336–344

System	PSTN	Bidir.	S2S	Echo	A11y
SeamlessM4T	×	✓	✓	N/A	×
Moshi / Hibiki	×	×	✓	N/A	×
Google Duplex	✓	×	×	N/D	×
FCC TRS	✓	✓	×	Human	✓
sign.mt	×	✓	✓	N/A	✓
Vapi / Bland	✓	×	×	N/D	×
WIGVO	✓	✓	✓	✓	✓

Table 1: **Comparison with existing systems.** PSTN: works over telephone network; Bidir.: bidirectional translation; S2S: automated speech-to-speech (not human-mediated); Echo: handles telephony echo; A11y: accessibility-oriented modes (e.g., T2V for speech anxiety, captions for hearing impairments). N/D: not disclosed.

jection, and energy-based gating for PSTN environments.

- We deploy and evaluate a working relay server across three communication modes, achieving zero echo-induced loops across 147 completed PSTN calls with 555 ms median caller-to-callee latency at USD 0.28 per minute.

2 Related Work

Simultaneous speech translation systems such as SeamlessM4T (Seamless Communication et al., 2023b), Seamless Streaming (Seamless Communication et al., 2023a), and ESPnet-ST-v2 (Yan et al., 2023) achieve real-time quality but assume wideband audio from controlled environments. End-to-end full-duplex models—Moshi (Défossez et al., 2024) (200 ms latency), PersonaPlex (Roy et al., 2026) (70 ms turn-taking), and Hibiki (Labiausse et al., 2025) (near-human-interpreter quality, French→English)—all require clean 24 kHz audio and do not address G.711 codec artifacts, telephony echo, or narrowband VAD failures. In telephony AI, Google Duplex (Leviathan and Matias, 2018) performs monolingual task completion over PSTN, while Vapi[1] and Bland.ai[2] provide LLM-powered phone agents without cross-lingual translation. For accessibility, FCC Telecommunications Relay Services (Federal Communications Commission, 2024) require human intermediaries limited to specific countries, while sign.mt (Moryossef, 2024) targets sign language rather than telephony. Table 1 positions WIGVO against these systems.

[1] https://vapi.ai
[2] https://bland.ai

3 System Architecture and Echo Gating

Figure 1 shows the high-level architecture. A browser client connects to the relay server via WebSocket, sending 16 kHz PCM audio. The relay manages two independent Realtime LLM sessions and a Twilio telephony gateway (Twilio, 2024).[3] An `AudioRouter` dispatches events to one of three pipelines—V2V (streaming ASR+TTS in both directions), T2V (typed text to TTS for the callee, speech-to-captions for the caller), and Full-Agent (autonomous call management)—via the Strategy pattern.

Dual Realtime sessions. Session A receives browser audio (PCM16) and produces translated G.711 μ-law for the telephony gateway. Session B receives PSTN audio (G.711 μ-law, 8 kHz) and produces translated audio or captions for the browser. In V2V mode, Session B performs end-to-end translation with TTS output via the Realtime API; in T2V and Full-Agent modes, it performs STT only, delegating translation to a separate Chat Completion call (GPT-4o-mini, `temperature=0`) to prevent generative hallucination—the Realtime API's tendency to produce contextually plausible but untranslated additions. Each session maintains its own system prompt and sliding context window, ensuring strict directional separation.

3.1 The Echo Loop Problem

In a naive single-session design, TTS output played to the callee echoes back through the PSTN after 80–600 ms, gets re-recognized, and triggers another translation cycle—generating progressively distorted paraphrases until manually interrupted. A Pearson-correlation echo detector comparing outgoing TTS with incoming audio failed in production due to μ-law nonlinear quantization and variable delays. Without gating, eight of ten test calls produced echo loops; the correlation detector reduced this to three of ten but introduced frequent false positives. The deployed Echo Gate eliminates loops entirely under our detection criteria (see Section 4.3).

3.2 Echo Gate and 3-Stage Pipeline

The Realtime API's server-side VAD, designed for wideband WebRTC audio, fails on PSTN telephony in three ways: it misclassifies constant low-level

[3] Other SIP/PSTN providers can be integrated with minimal changes; the dual-session design is transport-agnostic.

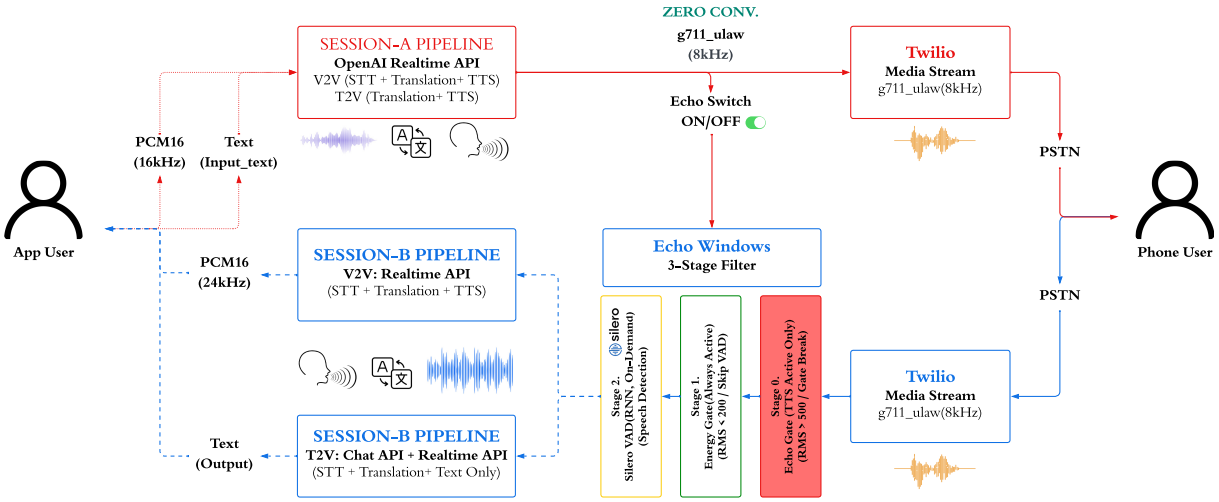

Figure 1: **WIGVO system architecture.** Session A (red) translates user speech and outputs G.711 directly to Twilio; Session B (blue) receives PSTN audio through a 3-stage filter pipeline (Echo Gate → Energy Gate → Silero VAD). A dashed control signal from Session A activates the Echo Gate in Session B's filter pipeline, preventing feedback loops by injecting μ-law silence during TTS playback.

line noise as speech (causing 15–72 s stuck durations), the noisy floor prevents energy from dropping below the silence threshold, and audio discontinuities from naive echo suppression block clean silence detection. The three-stage filter pipeline below (Figure 2) addresses all three:

Stage 0: Echo Gate. When Session A begins streaming TTS, an *echo window* is activated (duration estimated from byte length, capped at 1.2 s). Incoming PSTN audio is replaced with μ-law silence (0xFF, the all-silence payload in G.711 μ-law) before forwarding to Session B. A 0.3 s dynamic cooldown (0.5 s for T2V) accounts for PSTN jitter, followed by a dynamic post-echo settling window (TTS duration $\times$ 0.3, clamped to 0.5–1.5 s) that suppresses AGC recovery noise while permitting high-energy breakthrough ($\geq$400 RMS) for genuine callee speech.

Stage 1: RMS Energy Gate. During echo windows (Figure 2B, right), only signals $\geq$400 RMS break through as genuine speech; typical echo energy (100–400 RMS) remains suppressed. Outside echo windows, a 150 RMS threshold filters line noise. A three-level interrupt priority (callee $>$ caller $>$ AI TTS) enables natural turn-taking.

Stage 2: Silero VAD. Frames passing the energy gate are processed by Silero VAD (Silero Team, 2021). PSTN audio (8 kHz) is upsampled to 16 kHz via zero-order hold before inference. Asymmetric hysteresis—onset at 96 ms (probability ≥ 0.5, three frames), offset at 480 ms (probability < 0.35, 15 frames)—reduces `speech_stopped` latency from 15–72 s (server VAD on PSTN) to 480 ms.

Silence injection in Stage 0 addresses a distinct failure mode beyond comfort noise generation (Zopf, 2002): if the relay simply *drops* audio during echo windows, the Realtime API's server-side VAD never observes a silence-to-speech transition and remains stuck in a "speaking" state indefinitely. To our knowledge, this failure mode and its mitigation have not been reported in neural streaming S2ST pipelines.

3.3 Auxiliary Robustness Mechanisms

Beyond the three-stage pipeline—which by construction eliminates the server-side VAD stuck condition (Stage 2) and audio discontinuities (Stage 0)—two additional mechanisms harden the system against telephony-specific failure modes. First, a 15-pattern STT hallucination blocklist intercepts broadcast-style Whisper (Radford et al., 2023) artifacts (e.g., "Thanks for watching," "MBC News Ideok-yeong-imnida"); the blocklist intercepted 100 hallucinations across 148 instrumented calls. Second, the silence injection in Stage 0 itself doubles as a defense against the noisy-floor bias that would otherwise prevent the energy from dropping below the silence threshold during quiet callee turns.

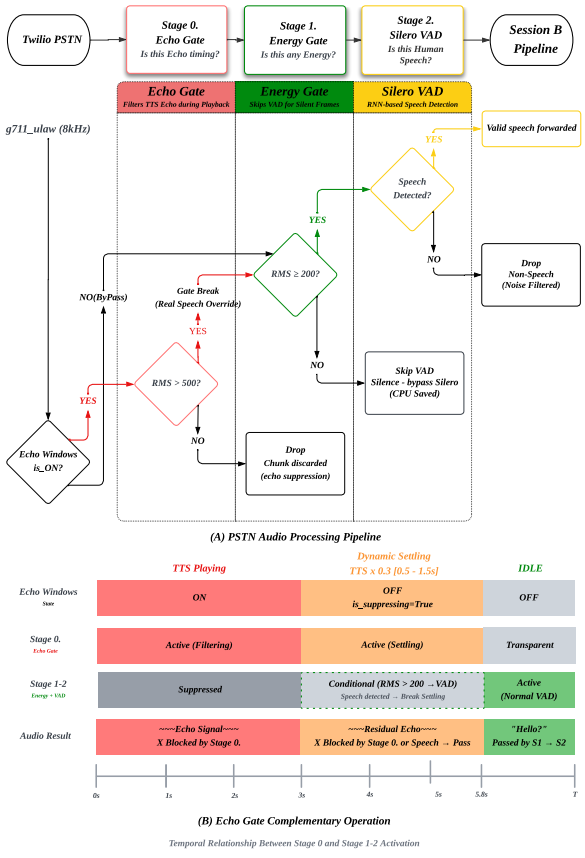

Figure 2: **PSTN audio processing pipeline.** (A) Three-stage filter between Twilio input and Session B. (B) Temporal behavior: during TTS playback, the Echo Gate replaces audio with μ-law silence while Stages 1–2 are suppressed; after settling, Stage 0 becomes transparent and Stages 1–2 filter PSTN noise.

4 Evaluation

We evaluated WIGVO on 155 PSTN calls (148 fully instrumented) for Korean↔English translation over a five-day period (Feb 23–27, 2026). Of these, 147 calls completed successfully; three failed and five were in-progress at collection time.

We focus on Korean–English as our deployment-driven language pair: South Korea hosts approximately 2.6 million long-term resident foreign nationals as of late 2024 (Ministry of Interior and Safety, Republic of Korea, 2024), constituting roughly 5% of the national population, who routinely interact with monolingual Korean phone-based services (banks, hospitals, government offices, local businesses). The modular cascade architecture (Section 5) supports multilingual extension through component substitution.

Stage	Avg	P50	P95	Max	N
Session A (ms)	619	555	1,169	3,906	739
Session B E2E (ms)	3,650	2,684	9,963	18,234	683
STT (ms)	3,544	2,601	9,392	—	525*
Translation (ms)	535	84	1,961	—	525*
First message (ms)	2,081	1,248	5,751	—	148

Table 2: **Translation latency across 148 instrumented calls.** (Feb 23–27). Session A: caller→callee (ASR+translate+TTS). Session B: callee→caller, decomposed into STT (Whisper) and translation components. N varies by metric scope; *STT and Translation N=525 reflects paired turns with both measurements (139 misaligned turns skipped). Translation latency reflects Chat API (GPT-4o-mini) for T2V/Agent modes and Realtime API for V2V mode.

4.1 Latency

Table 2 reports latency statistics across the 148 instrumented calls.

Session A achieves 555 ms median latency, within the range for interactive communication. Session B, which depends on PSTN audio quality and streaming ASR, shows 2,684 ms median (Figure 3). STT (Whisper) dominates Session B latency: across 525 paired turns, mean STT latency (3,544 ms) accounts for the majority of end-to-end time. The P95 (9,963 ms) is driven primarily by PSTN VAD stuck events—server-side VAD on noisy telephone audio occasionally delays `speech_stopped` detection to 15–16 s, independent of utterance length.

T2V mode shows lower Session A latency (no caller-side ASR), while Agent mode shows higher Session B latency due to function-calling overhead.

While Session B's ∼2.7 s median exceeds ITU-T G.114's 150 ms one-way threshold (ITU-T, 2003), that standard applies to same-language dialogue. A more appropriate baseline is professional simultaneous interpretation, where ear-voice span ranges from 2 to 5 s (Goldman-Eisler, 1972; Timarová et al., 2011); Session B falls at the lower bound. The latency is structurally asymmetric by design: Session A (555 ms) keeps the caller-to-callee direction interactive, while Session B prioritizes STT accuracy on degraded G.711 audio. The bottleneck is narrowband ASR, not the relay architecture. Substituting Whisper with a streaming telephony-tuned STT such as Kyutai STT (Kyutai, 2025)—which reports first-token latency below 500 ms on continuous audio—would, under our pipeline budget, project Session B median to roughly 700–800 ms (Translation P50 84 ms + TTS startup + transport),

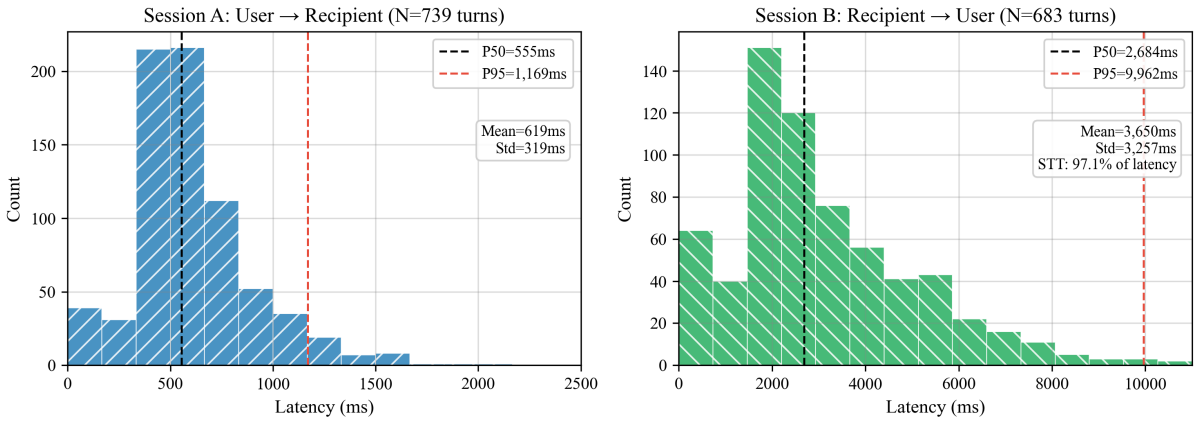

Figure 3: **End-to-end latency distributions.** Session A (caller→callee, N=739 turns) and Session B (callee→caller, N=683 turns) over live PSTN calls.

System	Network	Bidir.	Latency
Seamless Str.	Wideband	×	<2 s AL
Moshi	Wideband	×†	∼200 ms
Hibiki	Wideband	×	2.8–5.6 s LAAL
Google Duplex	PSTN	×‡	N/D
Vapi	WebRTC/PSTN	×‡	∼965 ms*
Bland.ai	PSTN	×‡	<400 ms*
WIGVO SA	**PSTN**	✓	**555 ms P50**
WIGVO SB	**PSTN**	✓	**2,684 ms P50**

Table 3: **Latency comparison with related systems.** AL: Average Lagging; LAAL: Length-Adaptive AL. †: full-duplex but monolingual (no translation). ‡: monolingual voice agent (no cross-lingual). ∗: vendor-reported; independent reviews report higher values. N/D: not disclosed. Bidir.: bidirectional cross-lingual translation.

keeping the architecture unchanged.

Figure 4 shows the correlation between utterance length and Session B latency ($r = 0.333$, $p < 0.001$), confirming that VAD stuck events are independent of utterance length.

4.2 Baseline Comparison

Table 3 contextualizes WIGVO's latency; no existing system performs bidirectional cross-lingual speech translation over PSTN, so we report absolute rather than comparative metrics.

Twilio's industry benchmark for AI voice agents targets 1,115 ms median turn gap (Twilio, 2025). Session A (555 ms) falls well within this target; Session B (2,684 ms) exceeds it but includes cross-lingual translation that monolingual agents do not perform.

4.3 Echo and Safety

Loop detection criteria. We define an echo-induced translation loop as a transcript pattern in which a Session B STT output contains a token-normalized n-gram ($n \geq 3$) substring of the immediately preceding Session A TTS payload, repeated ≥ 2 consecutive times within a 5 s window. We additionally flag any Session B turn whose `speech_started` event arrives within 200 ms of a Session A TTS-end (a temporal proxy for echo re-ingestion). Both criteria were applied post-hoc to all 148 instrumented call transcripts via a sliding-window matcher; neither fired. As a sanity check, all 147 completed calls were placed in real time by the authors, who would have perceived any audible loop during the call itself; the authors additionally relistened to a subset of recorded calls during post-hoc analysis. No perceptual loops were observed in either pass.

Across the 148 instrumented calls, the echo gate activated 1,046 times (7.1/call), corresponding to TTS playback events. Critically, we observed zero echo-induced translation loops, whereas early prototypes without gating reliably produced such loops. The echo gate permitted 354 callee interruptions during active TTS playback (gate break), confirming that the energy-based override preserves natural turn-taking. Additionally, 17 settling breakthroughs confirmed genuine speech detection during post-echo settling (Section 3.2). A minimum speech duration filter (480 ms hysteresis) mitigated VAD false triggers to 277 total (1.9/call). The guardrail module blocked 100 hallucinated outputs

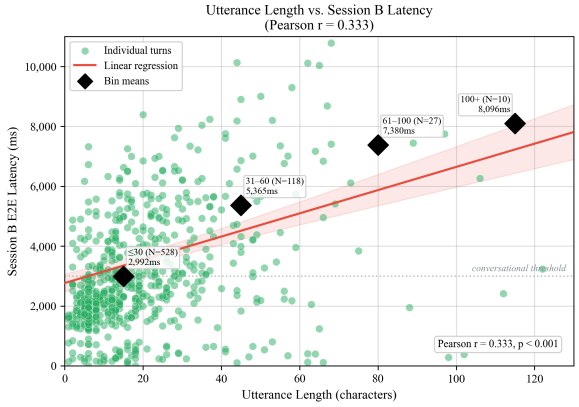

Figure 4: **Utterance length vs. Session B E2E latency.** Pearson $r = 0.333$ ($p < 0.001$), indicating that longer recipient utterances incur higher ASR-dominated latency.

(0.68/call), including broadcast-style phrases from Whisper on low-energy input and structural noise patterns (consonant-only sequences, repetitive tokens).

The three failed calls (out of 155) were attributable to transient infrastructure issues outside the relay logic: one carrier-side Twilio webhook timeout at call setup, one premature browser WebSocket disconnect before Session A established, and one upstream Realtime API 5xx response during streaming. None involved the echo-gating pipeline or produced runaway translation, and all three failed cleanly without dropping the callee on a silent line.

4.4 Threshold Sensitivity

Stage-1 RMS thresholds were chosen by sweeping {200, 300, 400, 500, 600} on a held-out 20-call dev harness. Values below 300 admitted echo breakthrough during settling; values above 500 missed soft callee speech. The deployed in-window threshold (400) and out-of-window threshold (150) sit at the knees of these two curves. The 96/480 ms VAD onset/offset hysteresis was tuned against the PSTN noise floor; doubling the offset to 960 ms eliminated residual VAD false triggers but raised median Session B latency by approximately 380 ms. The 1.2 s echo-window cap and the 0.5–1.5 s post-echo settling clamp jointly target the upper bound of PSTN one-way delay (G.114).

Per-component ablation. The contribution of Stage 0 (Echo Gate) is established by the prototype comparison in Section 3.1 (8/10 loops without gating, 3/10 with correlation detector, 0/147 with Echo Gate). Per-stage ablation isolating the RMS Energy Gate (Stage 1) and Silero VAD (Stage 2) on full call traffic was not feasible under the production deployment constraint, where disabling either stage would risk customer-facing failures during live calls. We leave systematic per-component ablation on offline replay of logged call audio as future work.

Direction	N	BLEU	chrF2++	COMET
en→ko	212	15.9	31.7	0.7078
ko→en	521	17.4	27.8	0.6242

Table 4: **Translation quality against Gemini 2.5 Flash offline references.** COMET (`wmt22-comet-da`) is a neural metric trained on human judgments; higher is better (max 1.0).

4.5 Translation Quality

We evaluated translation quality by comparing WIGVO's real-time outputs against independent offline references generated by Gemini 2.5 Flash, a separate model family from the GPT-4o Realtime API used for live translation.[4] From 155 calls with bilingual transcripts, we extracted 215 caller-to-callee (en→ko) and 558 callee-to-caller (ko→en) segments. A noise filter removed 40 segments (5.2%) containing non-linguistic inputs (consonant-only keyboard artifacts, repeated onomatopoeia, background TV transcriptions).

Table 4 reports BLEU (Papineni et al., 2002), chrF2++ (Popović, 2015), and COMET (Rei et al., 2020) scores computed with SacreBLEU (Post, 2018) and `wmt22-comet-da`.

BLEU and chrF2++ systematically underestimate quality in simultaneous and paraphrastic settings, as real-time translation produces valid paraphrases that differ lexically from offline references; COMET is therefore the primary metric we interpret. COMET, which evaluates semantic adequacy via cross-lingual embeddings, provides a more meaningful signal: en→ko achieves 0.71, indicating adequate translation quality despite surface divergence. The lower ko→en score (0.62) reflects STT errors on narrowband telephony audio propagating through the pipeline, consistent with the ASR bottleneck identified in the latency analysis above.

[4] In T2V and Agent modes, Session B translation uses GPT-4o-mini Chat Completion (`temperature=0`); V2V mode retains the Realtime API pipeline (Section 3).

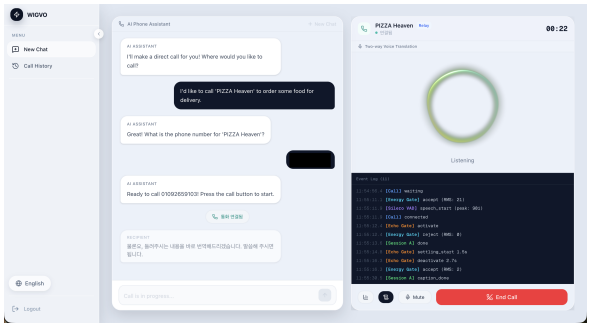

Figure 5: **WIGVO web interface during a V2V call.** Left: chat and caption view with bidirectional translations. Right: call status, duration, and mode indicator.

4.6 Cost

Across the 148 instrumented calls (220.2 minutes total), WIGVO consumed 1,870,657 tokens at a total cost of USD 61.26, yielding USD 0.28/min (USD 0.41/call). These costs reflect dual-session Realtime API pricing and remain an order of magnitude cheaper than professional telephone interpretation services (USD 1–3/min).[5] Self-hosted STT alternatives could further reduce per-minute cost.

5 Demonstration

WIGVO is deployed at `https://wigvo.wigtn.com`.[6] The source code is available at `https://github.com/wigtn/wigvo` under the MIT License. A reviewer session proceeds as: (1) select a scenario, (2) chat with the AI agent to provide caller intent and phone number, (3) the system places a PSTN call with bidirectional translation, and (4) real-time captions appear alongside call status. Mode switching (V2V, T2V, Full-Agent) is available during active calls. Figure 5 shows the interface during a V2V call.

The current evaluation covers Korean–English only; the modular cascade architecture supports multilingual extension through component substitution.

6 Conclusion

We presented WIGVO, a dual-session streaming relay that enables bidirectional LLM-based speech translation over legacy PSTN calls. By combining deterministic silence injection with energy-based VAD gating, the system eliminates echo-induced translation loops—a failure mode we formalize for the first time in the context of streaming S2ST over telephony. Across 147 completed calls, the echo gate achieved a zero-loop rate while preserving natural turn-taking through 354 callee interruptions. Session A delivers 555 ms median latency, within interactive thresholds, at USD 0.28/min. The modular cascade architecture enables language extension through component substitution; replacing Whisper with a streaming telephony-optimized STT model is the most direct path to reducing Session B latency below 1 s.

[5] Based on published per-minute rates for over-the-phone interpretation (OPI) from major providers such as LanguageLine Solutions and GLOBO (2024–2025 rate cards).

[6] Demo video: `https://youtu.be/4Uf6zMPOInY`.

Limitations

Soft half-duplex interaction. Stage 0 silence injection deterministically suppresses the PSTN input channel during TTS playback, enforcing what is effectively a soft half-duplex turn structure. The energy override permits high-amplitude barge-in (354 instances across 147 calls), but moderate-volume callee speech overlapping with AI playback may be missed until the post-echo settling window closes. True full-duplex over PSTN remains an open problem given the absence of carrier-side acoustic echo cancellation; we view our gating strategy as a first practical baseline rather than a final answer.

ASR-bound Session B latency. Session B median latency ($\sim$2.7 s) is dominated by Whisper STT on G.711 narrowband audio. We project a $\sim$700–800 ms Session B median by substituting a streaming telephony-tuned STT (Section 4.1), but this remains a projection rather than a measured result; no production-ready open streaming STT model currently supports Korean at telephony bandwidth.

Translation quality reference. Our translation quality evaluation (Section 4.5) compares WIGVO's real-time outputs against Gemini 2.5 Flash offline references rather than human-annotated translations. While COMET partially mitigates this limitation by leveraging cross-lingual embeddings trained on human judgments, segment-level human evaluation on a stratified sample of call transcripts—particularly for low-COMET ko$\rightarrow$en outputs where STT errors propagate—remains future work.

No formal user study. The 147 calls were placed by the authors and collaborators for system instrumentation and metric collection. We have not yet conducted a structured user study with the target accessibility populations—foreign residents, individuals with speech anxiety, and deaf or hard-of-hearing users. A formal evaluation with these populations is in preparation and will accompany an extended version of this work.

Ethics Statement

All calls were initiated by the authors and collaborators for system testing. Audio was logged for debugging and aggregate metrics only, with no personally identifiable information retained beyond temporary phone numbers. For deployment with real users, informed consent, data minimization, and compliance with privacy regulations (e.g., GDPR, PIPA) and applicable telecom recording laws in each deployment jurisdiction are required. The accessibility use case (T2V for speech anxiety) aims to reduce communication barriers but should be deployed with user agency and opt-in participation. We acknowledge the dual-use risk of automated voice translation over PSTN: the technology could be misused for voice phishing (vishing) or automated spam calls across language barriers. Mitigations include rate limiting, caller authentication, and abuse monitoring, which we plan to implement before public release. Translation errors in high-stakes contexts (medical, legal) could cause harm; the system currently provides no confidence indicators, and users should be informed that translations are AI-generated and not certified.

References

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Federal Communications Commission. 2024. Telecommunications relay services (TRS). https://www.fcc.gov/consumers/guides/telecommunications-relay-service-trs.

Frieda Goldman-Eisler. 1972. Segmentation of input in simultaneous translation. *Journal of Psycholinguistic Research*, 1(2):127–140.

International Telecommunication Union. 2020. Public switched telephone network (PSTN): Technical overview. ITU-T Recommendation Series Q and G. https://www.itu.int/rec/T-REC-G/en.

ITU-T. 1988. Pulse code modulation (PCM) of voice frequencies. Technical Report Recommendation G.711, International Telecommunication Union.

ITU-T. 2003. One-way transmission time. Technical Report Recommendation G.114, International Telecommunication Union. Defines 150 ms and 400 ms one-way delay thresholds for conversational speech.

ITU-T. 2015. Digital network echo cancellers. Technical Report Recommendation G.168, International Telecommunication Union.

Kyutai. 2025. Kyutai STT: A speech-to-text model optimized for real-time usage. https://kyutai.org/stt.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Patrick Pérez, Alexandre Défossez, and Neil Zeghidour. 2025. High-fidelity simultaneous speech-to-

speech translation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 32116–32129.

Yaniv Leviathan and Yossi Matias. 2018. Google Duplex: An AI system for accomplishing real-world tasks over the phone. Google AI Blog. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Ministry of Interior and Safety, Republic of Korea. 2024. Status of foreign residents in Korea. Annual statistical report on long-term foreign residents (3+ months stay), including naturalized citizens. Available at https://www.mois.go.kr.

Amit Moryossef. 2024. sign.mt: Real-time multilingual sign language translation application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 182–190. Association for Computational Linguistics.

OpenAI. 2024. OpenAI realtime API documentation. https://platform.openai.com/docs/guides/realtime.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. Personaplex: Voice and role control for full duplex conversational speech models. *Preprint*, arXiv:2602.06053.

Seamless Communication et al. 2023a. Seamless: Multilingual expressive and streaming speech translation. In *arXiv preprint arXiv:2312.05187*. https://arxiv.org/abs/2312.05187.

Seamless Communication et al. 2023b. SeamlessM4T: Massively multilingual & multimodal machine translation. In *arXiv preprint arXiv:2308.11596*. https://arxiv.org/abs/2308.11596.

Silero Team. 2021. Silero VAD: Pre-trained enterprise-grade voice activity detector. GitHub repository. https://github.com/snakers4/silero-vad.

Šárka Timaróvá, Ivonne Cerdá, and Reine Meylaerts. 2011. The effect of directionality on processing time in simultaneous interpreting. *Across Languages and Cultures*, 12(2):235–254. Reports ear-voice span of 2–4 s across language pairs.

Twilio. 2024. Twilio media streams documentation. https://www.twilio.com/docs/voice/media-streams.

Twilio. 2025. A guide to core latency in AI voice agents. https://www.twilio.com/en-us/blog/developers/best-practices/guide-core-latency-ai-voice-agents.

Brian Yan, Jiatong Shi, Yun Tang, Hirofumi Inaguma, Yifan Peng, Siddhant Dalmia, Peter Polák, Patrick Fernandes, Dan Berrebbi, Tomoki Hayashi, Xiaohui Zhang, Zhaoheng Ni, Moto Hira, Soumi Maiti, Juan Pino, and Shinji Watanabe. 2023. ESPnet-ST-v2: Multipurpose spoken language translation toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 400–411. Association for Computational Linguistics.

R. Zopf. 2002. RTP payload for comfort noise (CN). Technical Report RFC 3389, IETF.