

Comparing Machine and Human Simultaneous Interpreting

Dongpeng Pan and Kilian G. Seeber

Faculty of Translation and Interpreting (FTI), University of Geneva
dongpeng.pan@unige.ch; kilian.seeber@unige.ch

Abstract

We evaluate four machine simultaneous interpreting (SI) systems, an open-weight and a commercial instance of both the cascaded and the end-to-end speech-native architecture, on simulated conference discourse interpreted into five languages. All outputs are compared with professional interpretations. Reference-free quality metrics are first calibrated against blind expert ratings of the human renditions: their association with expert ratings is strongest for content, weaker for presentation, and weakest for style. Machine systems attain content-similarity scores within or above the professional range and higher median segment coverage at the reported alignment threshold. Coverage can count professional compression as omission and is sensitive to transcription and alignment error, so the surplus does not establish better interpreting. For latency, only the commercial speech-native system approaches professional timing. Estimated content lag thus ranges from under twice to more than ten times the professional median, and each system’s delay is associated with an observable operating property: commit granularity, synthesis queuing, or compute limits. Together, the rating-calibrated metrics and professional distribution provide a reference-free procedure for evaluating machine interpreting on unsegmented, full-length conference discourse.

1 Introduction

Simultaneous speech translation is deployed on long-form, unsegmented, spontaneous speech, whereas its established evaluation methods presuppose sentence-segmented input with reference translations (Ma et al., 2020; Papi et al., 2025). In simultaneous interpreting (SI), professional renditions can involve compression and syntactic restructuring (Liontoul, 2011; Hodzik et al., 2024). A corpus-based analysis of English–Japanese SI also describes reformulation through information packaging at conceptual and translation-unit levels

(Ishizuka, 2024). A single sentence-aligned reference may therefore fail to capture this variation, and expert evaluation panels cannot be convened as frequently as systems are updated. An evaluation matched to this setting must therefore (i) operate reference-free on unsegmented audio, (ii) apply identically to systems whose internals are inaccessible, and (iii) demonstrate, rather than assume, that its automatic metrics measure what expert judges of interpreting evaluate.

This paper contributes such an evaluation and applies it to the two architectures currently in use. Cascaded systems decompose the task into recognition, text translation, and synthesis; end-to-end speech-native systems translate speech directly without a committed text stage. We instantiate each family twice, as an open-weight system, whose internal states remain available for subsequent model analysis, and as a commercial service, which the protocol must handle as a black box. The systems are an open cascade (Whisper-large-v3, NLLB-200, MMS-TTS) and the commercial cascade Azure Speech Translation; the open end-to-end SeamlessStreaming and the commercial speech-native gpt-realtime-translate.

A completed experiment on distance interpreting (Seeber and Pan, 2026) provides renditions of the same four enacted conference scenarios (on average 10.5 minutes each) by 78 professional conference interpreters into six languages (74 enter the analyses; Section 3), with 3,016 discrete blind expert ratings on content, style, and presentation. The ratings validate the reference-free metrics before any machine output is scored (Section 6.1): the metrics are more strongly associated with the experts’ content judgments than with presentation or style, so machine claims are confined to content. The renditions define the *human performance envelope*, the empirical distribution of professional latency, coverage (the share of source segments receiving a rendition), and content-similarity scores

under the same instruments, in which every machine result is placed; the five machine languages and four speeches are fixed by this corpus.

Three research questions structure the study:

RQ1 Which reference-free quality metrics align with professional judgment of interpreting, on which rating criteria, and finely enough to resolve known quality differences?

RQ2 Where does each system score within the professional envelope on content similarity and content coverage?

RQ3 Does any system reproduce professional temporal behavior: the listener-experienced content lag and its regulation over a full speech?

Sections 3 and 4 describe the data and systems, Section 5 the measurement protocol, Section 6 the metric validation and the comparison.

2 Related Work

Evaluating simultaneous speech translation.

Latency evaluation for SimulST standardized around SimulEval (Ma et al., 2020) and the average-lagging family: AL (Ma et al., 2019), its length-adaptive correction LAAL (Papi et al., 2022), and DAL (Cherry and Foster, 2019). These measures presuppose sentence-segmented input, which deployment speech does not provide. StreamLAAL (Papi et al., 2024) extends latency measurement to unbounded streams, and Papi et al. (2025) show that reported SimulST performance degrades substantially under realistic long-form conditions. We provide measurement on unsegmented ecological speech with instruments that require no access to system internals, so that open models, closed services, and human interpreters are scored by the same procedure.

Reference-free quality estimation. Where suitable references are unavailable, quality estimation substitutes learned judgment: COMET-QE and CometKiwi on text (Rei et al., 2020, 2022), and, for speech, BLASER and BLASER-2.0 (Chen et al., 2023; Seamless Communication et al., 2023b), regressors over SONAR sentence embeddings (Duquenne et al., 2023) that score target speech directly against source speech. All are trained on translation-adequacy judgments; interpreting strategies such as compression and syntactic restructuring (Liontou, 2011; Hodzik et al., 2024) may complicate the interpretation of their scores. RQ1 tests whether, and for which criteria, these metrics align with expert interpreting judgment.

Human evaluation of machine interpreting.

Fantinuoli and Prandi (2021) framed machine-interpreting evaluation communicatively, arguing that MT metrics miss what users of interpretation need. Deployed systems have since been assessed by expert panels (Matsushita, 2026; World Health Organization, 2025), with reported lags up to 32 seconds and low acceptability rates. Such panels judge each machine anew, which cannot be repeated at the pace systems change. Our design places human judgment at the beginning instead: expert ratings that already exist, of human renditions, are used once to test which automatic metrics align with expert judgment (RQ1, Section 6.1), and the selected metrics can then score additional systems. Their validity on machine output still requires testing (Section 6.1). The two uses of human judgment are complementary.

An earlier report for ITCILO compared Wordly and Stenomatic with human interpreting on three event excerpts, using manually aligned transcripts and reference-free XCOMET-XL (ITCILO Advisory Services, 2025, pp. 14–19). The report cautioned that automated scores may penalize interpreters’ context-sensitive adaptations, including compression (pp. 37–38). This provides a methodological precedent for examining how automatic evaluation handles human interpreting.

Cascade and end-to-end systems. Both architecture families are currently in use, cascaded systems dominating the market and speech-native systems commercially available since 2026 (Adelani et al., 2026; OpenAI, 2026). A current comparison must therefore cover both in open-weight and commercial form. Our comparison anchors both architecture families in a distribution of professional performance on identical ecological input, with automatic metrics calibrated against expert ratings of human renditions.

3 The Human Benchmark

The human reference data originate from a completed experiment on distance interpreting (Seeber and Pan, 2026), conducted before any system described in this paper existed, and are reused here unchanged as measurement infrastructure. In that experiment, 78 professional conference interpreters each rendered the same four re-enacted conference scenes in English (roughly 10.5 minutes each; 2,559 s of source speech) in a counterbalanced 2×2 within-participant design crossing modal-

ity (audio-only, ARI, versus video remote, VRI) with sound quality (full-band HQ, 125 Hz–15 kHz, versus telephone-band LQ, 300 Hz–3.4 kHz). Target languages were French (45 interpreters), Spanish (11), German (6), Russian (6), Italian (5), and Dutch (5; no machine counterpart, retained only in envelope descriptives). The machine systems consume clean full-band audio, so the machine-matched human stratum is the audio-only, full-band condition (ARI-HQ). Machine placements are also checked against the remaining strata as a sensitivity analysis. Four of the 78 recorded participants are excluded on the original study’s pre-existing grounds (one produced no usable audio, one was dropped from the original quality evaluation, one completed only two of the four trials, one withdrew), leaving 74 in the envelope (App. A).

The same experiment produced a blind expert quality evaluation of the human renditions: professional raters (6 for French, 4 for Spanish, 5 each for German, Italian, and Russian) scored 8 fixed excerpt windows per interpreter on 7-point scales for content, style, and presentation, yielding 3,016 ratings of 69 interpreters with no missing cells (ICC = .642). These ratings provide the criterion for validating the reference-free metrics (RQ1, Section 6.1). No machine output is rated by humans, no new ratings are collected, and no individual interpreter is scored, ranked, or identified.

Ethics. The original experiment was conducted in line with the Declaration of Helsinki and approved by the Board of Ethics of the Faculty of Translation and Interpreting (FTI), University of Geneva; all participants gave informed consent. The present study is a secondary analysis of those recordings and ratings within the scope of that consent: no new human data were collected, rendition audio was processed exclusively as research data, and no individual interpreter’s results appear in this paper or its released artifacts.

4 Four Machine Systems

Table 1 summarizes four systems: one open-weight and one commercial system for each of two architectures, cascaded and end-to-end speech-native. Because the systems also differ in their models, training data, and implementation, the comparison cannot establish whether architecture or openness accounts for their performance differences.

Open cascade (Open-Cascade). A pipeline of open components: Whisper-large-v3 for recogni-

tion (Radford et al., 2023), NLLB-200-1.3B for translation (NLLB Team et al., 2022), and MMS-TTS for synthesis (Pratap et al., 2024), with a Piper voice for Italian because the MMS Italian checkpoint was withdrawn from distribution during the study (App. A). Whisper-large-v3 provides an openly available recogniser that can be run locally with fixed weights and decoding settings. The same checkpoint was used for English recognition in CUNI’s IWSLT 2025 simultaneous translation system (Macháček and Polák, 2025). Whisper’s translation mode produces English text; the present English-to-other-language directions therefore require a separate translation model (OpenAI, 2022). A streaming voice-activity detector commits speech to the pipeline: an utterance closes after 500 ms of silence, and an open buffer is force-committed at 8 s. Each chunk contains up to 8 s of source speech. Recognition, translation, speech synthesis, and waiting for playback can add further delay. The saved run records do not contain separate ASR start and finish times, so the ASR share of that delay was not measured.

Commercial cascade (Azure Speech Translation). Azure’s speech-translation service; the cascaded architecture is documented by the vendor (Microsoft, 2026). Synthesis is emitted per finalized recognition segment, a commit-policy property shared by deployed cascade products.

Open end-to-end (SeamlessStreaming). SeamlessStreaming (Seamless Communication et al., 2023b,a) translates speech to speech directly, with no committed text stage. The evaluated system is the streaming RunPod run on a rented A100. The model determines when to emit speech under its streaming policy. Machine content lag is reconstructed from client-side timing records under the playback model described in Section 5.

Commercial speech-native API (gpt-realtime-translate). A commercially available realtime speech-translation API, generally available since May 2026 at \$0.034 per audio minute (OpenAI, 2026). We classify it as end-to-end from vendor documentation; nothing in the comparison depends on the label.

Operating-point selection. The documented development sweeps used the first speech into French to select the lowest-latency setting that did not audibly damage output integrity. Azure Speech Translation kept its default semantic segmentation (the silence-timeout alternative halved com-

System	Class (as labeled)	Components / access	Cost
Open-Cascade	open cascade	Whisper-large-v3 → NLLB-200-1.3B → MMS-TTS (Piper for Italian); VAD commits: 500 ms close, 8 s cap; local GPU	≈9 GPU-h
Azure Speech Translation	closed cascade (vendor-documented ASR → TrueText → MT → TTS)	Speech SDK TranslationRecognizer; neural-voice synthesis; accessed 2026-08-03–04	\$0 billed; list ≈\$18–36
SeamlessStreaming	open end-to-end	SeamlessStreaming; client-side timing records; rented A100 (RunPod)	≈30 GPU-h
gpt-realtime-translate	closed; end-to-end per vendor documentation alone	realtime WebSocket API; accessed 2026-08-03	\$0.034/min

Table 1: The four machine systems: both architecture families in current use, each in open-weight and commercial form. Closed-system class labels are vendor-attributed and hedged (see text).

mit lag but truncated mid-clause); gpt-realtime-translate exposes no latency-relevant settings; Open-Cascade selected 500 ms/8 s. Settings and development sweeps are reported in Appendix A.

Unified input and timing protocol. The same source audio is pushed to each system as real-time-paced raw audio samples (PCM); client-side monotonic timestamps are recorded for every received audio chunk and text event; latencies are reported as distributions; each system’s quantization floor (buffer size, segment-finalization granularity, delta cadence) is disclosed alongside. All commercial runs occurred on August 3–4, 2026, to bound version drift.

5 Measurement Protocol

The protocol measures at two layers: a *comparison layer* (Layer A), for all human renditions and machine runs, and a *diagnostic layer* (Layer B), computed from machine client-side event streams. Figure 1 summarizes how recordings, transcription, alignment, expert-rating validation and system comparison connect.

Layer A (comparison): each full passage is transcribed (faster-whisper large-v3), its sentences aligned to the canonical source segments by multilingual-embedding similarity under a monotonic one-to-one dynamic program, and scored for content lag (onset-to-onset on the source clock) and content coverage. Matched source and target excerpts are scored using BLASER-2.0-QE on audio and SONAR text similarity on transcripts.

Coverage is the share of the speech’s canonical source segments (460 across the four speeches, 108–121 each) that receive a retained alignment to a target sentence; source segments without a retained match at or above the similarity floor of

0.25 are counted as uncovered. Because each target sentence can match only one source segment, a sentence combining content from several source segments cannot give coverage credit to all of them (App. A). The contributions of compression, restructuring and transcription error to the human–machine coverage difference have not been quantified.

Transcription is needed despite the reference-free metrics: it supplies sentence units for alignment and text for SONAR; audio-based BLASER needs no transcript. Machine transcript times are mapped from concatenated output audio to estimated playback positions using client-side timestamps and a sequential playback model. Without this remapping, machine content could appear to precede its source. The reconstructed machine lag has not been independently validated against recorded live playback. Client-side records also support the selection of machine excerpts.

Layer B (diagnostics): the machine event streams provide time to first audio, output delay relative to each system’s source timestamps, output tail, and real-time factor, where the required records are available. Expert ratings are available only for human excerpts and are used to assess the association of BLASER and SONAR scores with expert judgment (Section 6.1). Every human–machine comparison uses Layer A.

Alignment quality was hand-validated on complete side-by-side walkthroughs of three renditions per language before any downstream number was computed. Throughout, the professional envelope denotes the empirical distribution of per-rendition values in the machine-matched human stratum of Section 3; figures that display intervals show the 5th to 95th percentile of that distribution.

Every machine system completed all four pas-

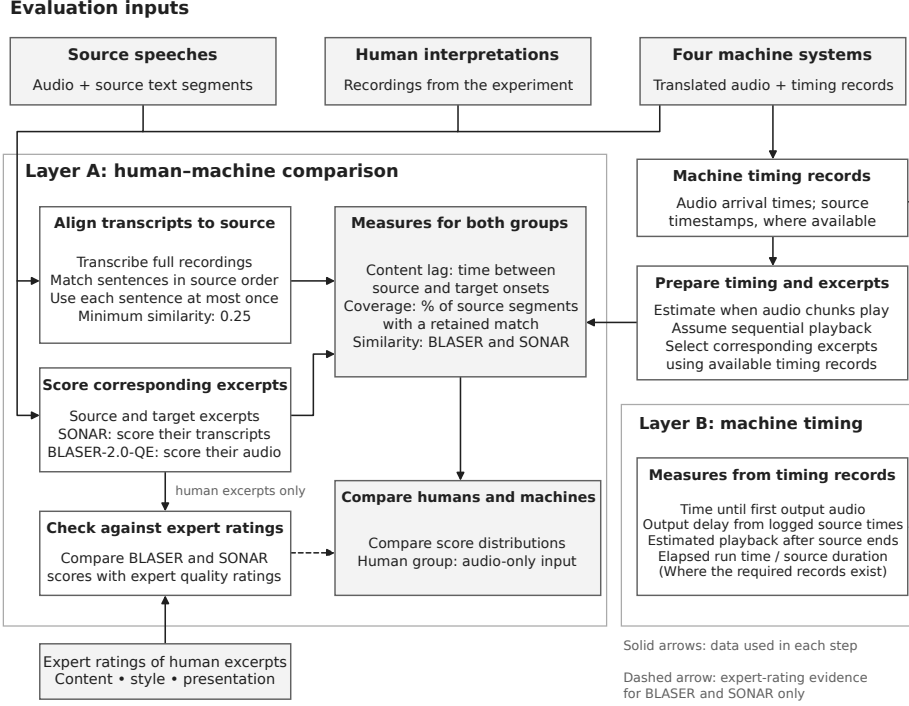


Figure 1: Evaluation workflow. Layer A compares human and machine interpretations using content lag, coverage and excerpt similarity. Expert ratings of human excerpts provide validation evidence for BLASER and SONAR (dashed arrow). Layer B reports timing measures derived from machine records.

sages in all five machine languages (20 runs per system). The archived alignment files contain 286 human renditions, densest in French (176) with 20 to 44 per other language. The 20 runs per system are not independent samples; every cross-system comparison is therefore paired within speech $\times$ language cells, and inference is conditional on these materials.

6 Results

6.1 Validity of the Similarity Metrics (RQ1)

Before BLASER and SONAR are applied to machine output, their relation to expert judgment is estimated on the excerpts the experts rated. The 3,016 blind ratings were matched to their excerpts, and each criterion was regressed on each metric in a cumulative-logit mixed model (ordinal::clmm, Laplace approximation, flexible thresholds) with random intercepts for rater, interpreter, and excerpt and language fixed effects. The excerpt intercept accounts for ratings of the same excerpt: each metric value attaches to one of 552 excerpts. Content was designated the primary criterion in advance; the remaining slopes are reported descriptively.

Both metrics showed their strongest estimated association with content ratings, followed by presentation and then style (Fig. 2): one SD of SONAR

text similarity multiplies the odds of a higher *content* rating by 1.66 (95% CI 1.48–1.87, likelihood-ratio $p < 10^{-16}$), a higher *presentation* rating by 1.26, and a higher *style* rating by 1.11 ($p = .017$); the audio-based BLASER-2.0-QE shows the same shape (1.26, 1.15, and a style odds ratio of 1.04, $p = .37$). The content slope is stable under leave-one-language-out refits (OR 1.53–2.05, significant in every refit); excerpt-level correlations are consistent (content .40 vs. style .15).

Machine similarity claims below are interpreted in relation to the content criterion; the study does not provide validated machine measures of style or presentation. The metric also resolves a known quality contrast: human raters registered a decrement of 0.13–0.22 points under degraded telephone-band sound, and the primary metric separates the same conditions ($\Delta = .031$, 95% CI [.005, .057]); the contrast serves only to test metric resolution, as the machine comparison uses full-band audio. Finally, the two reference-free metrics agree well across the 552 rated excerpts ($\rho = .70$): the two encoders give correlated rankings of the human renditions. We use SONAR as the primary similarity measure because its estimated association with expert content ratings is stronger.

This validity is established on human renditions; its transfer to machine output is an explicit as-

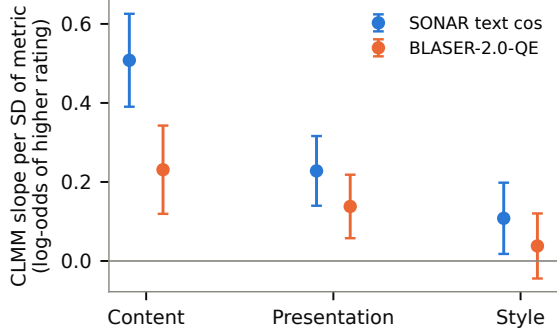


Figure 2: Criterion validation against 3,016 blind expert ratings: cumulative-logit mixed-model slope of each rating criterion on each metric (rating $\sim$ metric + Language + (1|Rater) + (1|interpreter) + (1|excerpt); error bars 95% CI). Associations are strongest for content, weaker for presentation, and weakest for style (BLASER: $p = .37$). This comparison supports the use of similarity as a content-related measure.

sumption, not a demonstrated fact. Fluent hallucinated content can escape a similarity metric (Section 6.5), and machine scores above the human range are extrapolations on the validated axis. Machine placements are therefore reported as *similarity* percentiles within the human distribution, never as expert-rated quality.

6.2 Placement in the Human Envelope (RQ2)

Figure 3 shows content lag on the horizontal axis, against coverage (Section 5) on the vertical. In the machine-matched ARI-HQ stratum, interpreters have a median content lag of 3.2 s and median measured coverage of 75%, with unmatched segments potentially reflecting compression, restructuring, actual omissions, or transcription and alignment error (Section 6.5). On these materials every machine system has measured coverage at or above the professional median (medians 83%–94% vs. human 75%); coverage, however, rewards literal retention (Section 6.4) and is therefore not a claim of better interpreting. Only gpt-realtime-translate delivers its coverage in human time.

On the calibrated similarity axis (Section 6.1), gpt-realtime-translate’s window similarity sits at the 90th–100th percentile of the human similarity distribution and Azure Speech Translation’s at the 100th; placements above the human maximum are extrapolations. Percentiles are capped at the 100th percentile and therefore cannot separate values above the human maximum. Equal percentile placements do not imply equal raw similarity; Table 2 reports raw similarity separately by language.

6.3 Temporal Behavior (RQ3)

The four systems have median content lags (gpt-realtime-translate 5.3 s, Azure Speech Translation 25.7 s, SeamlessStreaming 39.9 s, OpenCascade 40.5 s) against the human 3.2 s (Fig. 5). SeamlessStreaming was evaluated in its streaming RunPod configuration on a rented A100; its content-lag values are playback-model estimates (App. A). Distributional distances from the ARI-HQ human lag distribution ($n = 71$ per-rendition medians) are large for every system except gpt-realtime-translate: the 1-Wasserstein distance, the average shift in seconds needed to match a system’s lag distribution to the human one, is 2.5 s for gpt-realtime-translate, 25.6 s for Azure Speech Translation, 48.1 s for SeamlessStreaming, and 33.5 s for OpenCascade; the Kolmogorov-Smirnov distance, the largest gap between the two cumulative distributions, is ≥ 0.94 for every system. Only gpt-realtime-translate overlaps the human band and sits systematically on the slow shoulder of the human distribution.

The event streams provide system-specific timing diagnostics: Azure Speech Translation commits quickly once its recognizer closes a segment (median 1.3 s commit-to-audio) but waits for semantic segment boundaries (~ 7 s units) and then queues synthesized audio behind earlier playback; OpenCascade’s median per-commit processing delay is 2.6 s (the 8 s cap limits source buffering, not processing or playback delay) but its synthesis speaks at nearly the source’s pace, so playback debt compounds across the passage; SeamlessStreaming’s timing depends on the model’s emission decisions and computation time (App. A). These processes may contribute to the wide per-run spread of the three slower systems (IQRs spanning tens of seconds): lag accumulates with each speech’s density and pause structure. gpt-realtime-translate, which streams continuous ~ 200 ms audio deltas instead of queuing synthesized segments, is the only system whose per-run lags concentrate near a stable value. Human per-rendition median lags have a narrow interquartile range (2.6–3.5 s) across these 10.5-minute passages. This between-rendition pattern does not directly establish how lag is regulated within a passage.

On these materials, the fastest system combines relatively low content lag with high measured segment coverage and source similarity. This comparison does not establish a causal quality–

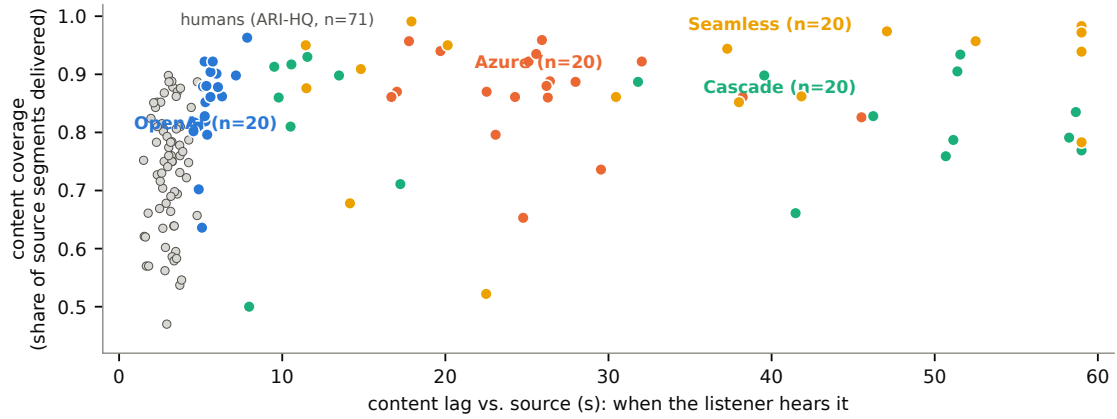


Figure 3: Content lag versus content coverage, per rendition/run. Each gray point is one professional interpreter’s full-passage rendition (ARI-HQ stratum); colored points are machine runs (one per speech×language). Content lag is onset-to-onset on the source clock: for humans, directly from the alignment; for machines, transcript times are first mapped from concatenated output audio to their queue-model playback position. Lags >59 s drawn at the frame edge.

	fr	de	es	it	ru
Humans (ARI-HQ)	0.67	0.57	0.67	0.58	0.62
gpt-realtime-translate	0.75	0.74	0.73	0.72	0.65
Azure Speech Translation	0.82	0.82	0.85	0.86	0.83
SeamlessStreaming	0.65	0.65	0.76	0.71	0.63
Open-Cascade	0.76	0.73	0.80	— ^a	0.74

Table 2: Per-language content similarity to source windows (SONAR text cosine, median over rated windows); the human row is the machine-matched stratum (audio-only, full-band sound; ARI-HQ). Most machine-language medians meet or exceed the corresponding human median; source similarity can penalize professional compression and does not establish interpreting quality. ^aItalian Open-Cascade windows were not scored (runs completed post-freeze, Section 4); their run-level lag and coverage are included in all analyses.

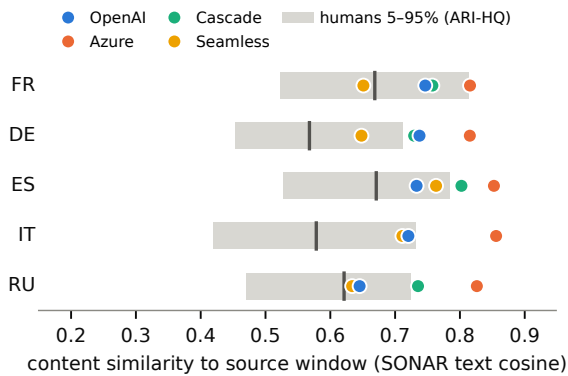


Figure 4: Human SONAR-similarity distributions per language (5–95%, ARI-HQ) with machine medians.

latency trade-off. Paired coverage differences between systems span 1 to 6 points across the 20 speech×language cells (cell-level paired differences resampled with replacement, 2,000 replicates, percentile intervals), with intervals crossing zero for every pair, so the data establish neither an ordering nor equivalence, and both commercial systems place at the envelope top despite a five-fold lag difference.

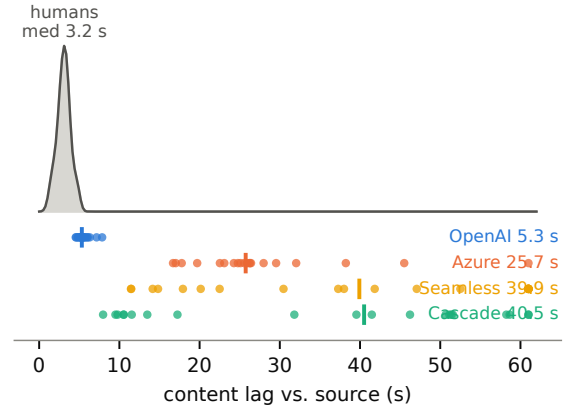


Figure 5: Content lag on the source clock: human density (per- rendition medians, machine-matched ARI-HQ stratum, $n = 71$) with per-run machine values. Machine transcript times remapped to the source clock (App. A).

6.4 Stability, Cost, and the Human Scale

Figure 4 illustrates that higher source similarity can reflect greater literal retention. The measure does not distinguish professional compression from content loss or measurement error, so its values do not constitute an overall quality ranking. Costs: gpt-

	n runs	content lag	coverage	first audio	tail	similarity pctl
Humans (ARI-HQ)	71 ^a	3.2 s [2.6–3.5 s]	75%	–	–	(envelope)
gpt-realtime-translate	20	5.3 s	86%	0.8 s	6.5 s	90th–100th
SeamlessStreaming	20	39.9 s ^c	94%	5.1 s	56.5 s	96th (IQR 33–100)
Azure Speech Translation	20	25.7 s	88%	13.8 s	30.2 s	100th
Open-Cascade	20 ^b	40.5 s	83%	12.4 s	54.2 s	100th (IQR 45–100)

Table 3: Main comparison. Lag, coverage and timing entries summarize runs or renditions; similarity percentiles summarize excerpt scores. Content lag = source-to-target onset delay on the source clock, reconstructed for machines; coverage = share of source segments with a retained alignment; similarity pctl = placement of window SONAR similarity in the ARI-HQ human distribution (the content-anchored proxy of Section 6.1; values above the human range are extrapolations). ^arenditions in the machine-matched stratum (audio-only, full-band sound; ARI-HQ), of 286 aligned across all strata. ^bItalian runs completed post-freeze via the repaired Piper voice and are included (App. A). ^cstreaming run on a rented A100 (RunPod).

Source (opening): “Without further ado, let’s get started. Firstly, I’d like to welcome you all to what promises to be a really interesting discussion...”
SeamlessStreaming: “*Sans plus de douille*, on commence. Premièrement, j’aimerais vous accueillir dans ce que promet d’être *des sites intéressants*...”
Source (final words): “... Well, firstly, thank you for having me, of course.” [interview continues to cutoff]
gpt-realtime-translate: “... Je vous remercie d’avoir écouté *cette vidéo*... j’espère que vous aurez apprécié cette vidéo. Merci à tous. À bientôt.”

Figure 6: Orthogonal failure modes (both ARI-4, French). Top: SeamlessStreaming commits before disambiguating context arrives; “without further ado” becomes the nonsensical *sans plus de douille* (“douille” = cartridge casing). Bottom: gpt-realtime-translate invents a complete video-outro that has no counterpart in the source when the input trails off. The output is fluent, polite, and entirely hallucinated.

realtime-translate cost \$7.25 at \$0.034/min; Azure Speech Translation \$0 billed (free-tier credits; list-price equivalent $\approx$ \$18–36 for both sweeps); the open systems cost GPU time ($\approx$ 9 GPU-h, $\approx$ 30 GPU-h), SeamlessStreaming requiring datacenter-class compute to approach real time (Section 6.3).

6.5 Qualitative Failure Modes

SeamlessStreaming commits *too early* on partial context, which surfaces as broken lexical choices an expert would score under *content*. gpt-realtime-translate *hallucinates*: it completes *too fluently* at the end of input, producing confident invented content that a similarity metric barely penalizes (its full-passage SONAR cosine is *higher* than SeamlessStreaming’s, 0.85 vs. 0.78). Neither similarity placement nor criterion calibration reliably detects the hallucinated outro; only manual source-output inspection revealed it. Both failures were found by listening to runs and inspecting transcripts against the source during alignment vali-

dation; no exhaustive hallucination census across all eighty runs was performed, so the frequency of such events is unknown (Limitations).

7 Discussion

Timing differences at the tested operating points.

The SimulST literature treats latency as a controllable policy parameter: wait- k policies trade quality against a chosen delay (Ma et al., 2019), learned read-write policies make that trade adaptive (Seamless Communication et al., 2023b), and the average-lagging metric family scores exactly this trade (Ma et al., 2019; Papi et al., 2022). All paired bootstrap intervals for coverage cross zero, establishing neither an ordering nor equivalence (Section 6.3). The commercial systems differ approximately five-fold in median lag while both occupy the upper end of the human similarity distribution; the capped percentiles cannot quantify their similarity difference. The timing records identify candidate contributors to these differences: semantic-segment closure and synthesis queuing for the cascades, computational demands for the open speech-native model, and continuous $\sim$ 200 ms emission for the commercial speech-native system. Cherry and Foster (2019) argued that average-lagging values are not comparable across systems with different commitment behavior; at passage scale, the comparable quantity is the listener-experienced lag distribution on a shared source clock.

Computation and passage-scale latency. Papi et al. (2025) showed that reported SimulST performance degrades substantially once computation and realistic long-form input enter the clock. For speech-to-speech systems, content-lag estimates depend on both output-emission decisions and the computation required on the reported hardware (Section 6.3). Latency claims must there-

fore report hardware and compute budgets. Our content-lag measure extends the resegmentation move of StreamLAAL (Papi et al., 2024) to speech-to-speech with no committed text on either side, allowing one instrument to score humans, open models, and closed services.

Hallucination and over-generation at passage scale. LAAL exists because over-generation gamed average lagging at sentence scale (Papi et al., 2022). Two passage-scale analogues appear here. First, machine coverage near 90% against the professional 75% does not establish superior interpreting. Greater literal retention is one possible explanation, but coverage alone cannot distinguish useful compression, substantive omissions, and transcription or alignment error. Second, the invented outro (Fig. 6) is over-generation at the input boundary, and it exposes the metric consequence: embedding similarity (Duquenne et al., 2023; Chen et al., 2023) rewards fluent topical continuation, so the hallucinated ending outscores a faithful but fragmented alternative. Quality-estimation metrics inherit the norms of the judgments they are trained on (Rei et al., 2020, 2022), and those norms are translation adequacy, not interpreting: they read strategy as loss and invention as adequacy. This supports delivery-aware, multimodal evaluation with access to prosody, seams, and timing. For presentation, candidate indicators include pause distributions, articulation rate, pitch variation, and audible repetitions or synthesis boundaries, measured from delivered audio. For style, transcript-based assessments could examine register and idiomatic expression. These indicators would require validation against the corresponding expert ratings, followed by testing on rated machine excerpts; none is treated here as an established substitute for those judgments. The human reference has a median lag of 3.2 s (IQR 2.6–3.5 s across rendition medians). Audible boundaries in segment-committed synthesis warrant direct assessment under presentation.

What would move each architecture into the envelope. The timing diagnostics (Section 6.3) suggest engineering changes to test. The cascades’ latency floor is set by segment closure plus synthesis at speech pace; newer cascades with an LLM translation stage (Grigoryan et al., 2026) warrant comparison, but this study does not measure their similarity or latency. Time-aware synthesis (TTS that accepts a duration target) and sub-segment commitment are the levers. For open speech-native

systems, emission policy and computational capacity should be evaluated together. That the only system inside the human band streams continuous audio deltas suggests commit granularity, not architecture family, is the operative variable, a hypothesis the next system generation can test. Expert panels judging deployed systems (World Health Organization, 2025; Matsushita, 2026) reported the same tens-of-seconds lag scale our instruments measured; the one in-band system shows such verdicts are policy-specific, not generic.

A recipe for human-anchored evaluation of machine interpreting. Applying the protocol elsewhere involves five steps: (1) reuse or obtain a rated human corpus for the register; (2) validate candidate reference-free metrics per rating criterion against its ratings, keeping only the criteria they align with; (3) run every system on the same passages as real-time-paced audio, logging client-side events; (4) apply the scoring protocol to humans and machines and quantify transcription and timing-estimator differences; (5) report placements in the human distribution with paired contrasts. Human calibration can be reused within a corpus, but transfer to new machine outputs requires checks for system-specific failure modes.

Future work. The corpus supports three extensions: blind expert ratings of stratified machine excerpts to test transfer of criterion validity; listener studies of audience comprehension and listening effort under matched conditions; and mechanistic analysis of the open systems, whose attention states are observable during inference.

8 Conclusion

We compared four machine simultaneous-interpreting systems with the professional reference distribution: human renditions of the same discourse, with automatic metrics assessed against 3,016 blind expert ratings of 552 excerpts from 69 interpreters. The metrics were most strongly associated with content judgments. Machine outputs attained similarity scores within or above the human range and higher measured segment coverage, but all four systems had longer median content lag; only gpt-realtime-translate approached professional timing. Higher coverage does not establish better interpreting, because the measure can penalize compression and is sensitive to transcription and alignment error. The protocol supports reproducible comparison against a profes-

sional reference, while the validity of its similarity measures on machine output remains to be tested with expert ratings.

Limitations

Vendor and architecture confounds. The commercial pair differs in vendor, training data, voices, and endpointing. That contrast is descriptive only; no architecture or openness main effect is claimed. Closed-system architecture labels rest on vendor documentation: gpt-realtime-translate’s transcript streams mean an internal text stage cannot be excluded from outside; systems can instead be grouped by observable commit behavior (continuous audio deltas versus segment-finalized synthesis) without changing measurement. Both commercial arms are dated snapshots (August 3–4, 2026; model strings in App. A) and may not replicate after vendor updates.

Hardware dependence. The evaluated natural SeamlessStreaming rerun used a rented A100. Its reported latency characterizes that configuration and hardware and is not a lower bound for the architecture. Open-Cascade’s Italian runs completed after the operating-point freeze via a repaired TTS voice and are included (App. A). Machine content lag is reconstructed from client-side timestamps under a sequential playback model and has not been independently validated against recorded live playback.

Materials and cohort. Four speeches, one register (prepared conference roundtable), one source language (English), five target languages; the 20 runs per system are those four speeches crossed with five languages, so all conclusions are conditional on these materials and none generalizes beyond this register without new data. The envelope is densest in French (45 interpreters vs. 5–11 elsewhere), so per-language placements outside French carry wide intervals. ASR errors may differ between live interpreter speech and synthesized machine speech, affecting both text similarity and alignment-based coverage. The audio metric and alignment hand-validation do not quantify or correct this possible asymmetry. Restructured renditions can also register spurious omissions. The contribution of transcription and alignment error to the observed human–machine gap remains unquantified.

Transfer of criterion validity. Expert ratings cover human renditions only; no machine output

was rated. Criterion validity is therefore established on human speech and *transferred* to machine output, and machine scores above the human range are extrapolation on the validated axis. The invented outro of Section 6.5 shows the failure concretely: fluent hallucination is barely penalized by the proxy, and no systematic hallucination census across the eighty runs was performed, so the frequency of such events is unknown. The machines’ fragment prosody and seam repetitions fall precisely on the unmeasured style and presentation criteria. Blind expert ratings of a stratified sample of machine windows are needed to test transfer. The open cascade is the reproducible baseline specified in Section 4. Its results characterize that configuration and hardware and do not establish the performance of more recent or faster cascades.

References

- David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sébastien Bratières, Marine Carpuat, Fabrício Carraro, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, Lewis C. Howe, Dávid Javorský, Yejin Jeon, and 36 others. 2026. [Speech translation and metrics in 2026: Findings of the IWSLT campaign](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 336–422, San Diego, USA (in-person and online). Association for Computational Linguistics.
- Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.
- Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.
- Claudio Fantinuoli and Bianca Prandi. 2021. Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 245–254, Bangkok, Thailand (online). Association for Computational Linguistics.

- Lilit Grigoryan, Vladimir Bataev, Andrei Andrusenko, Oleksii Hrinchuk, Davit Karamyan, Enas Albasiri, Vitaly Lavrukhin, Nikolay Karpov, and Boris Ginsburg. 2026. [NeMo@IWSLT 2026: Cascaded system for simultaneous speech translation](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 204–211, San Diego, USA (in-person and online). Association for Computational Linguistics. Cascade of a dual-mode ASR transducer and a multilingual LLM for the IWSLT 2026 simultaneous tracks.
- Ena Hodzik, Semra Özdemir, Nesrin Conker, and Orhan Bilgin. 2024. Strategic syntactic restructuring during simultaneous interpreting from Turkish into English. *Meta*, 69(1):242–263. <https://doi.org/10.7202/1113949ar>.
- Hirofumi Ishizuka. 2024. Two levels of information packaging and cognitive operations during simultaneous interpreting: An analysis via additional demonstratives. *Ampersand*, 12:100165. <https://doi.org/10.1016/j.amper.2024.100165>.
- ITCILO Advisory Services: *Comparison of the Technical Aspects of Live Translation Platforms during ITCILO Activities*. 2025. Final report, 3 February. Report commissioned by the International Training Centre of the International Labour Organization.
- Konstantina Lontou. 2011. Strategies in German-to-Greek simultaneous interpreting: A corpus-based approach. *Gramma: Journal of Theory and Criticism*, 19:37–56. <https://doi.org/10.26262/gramma.v19i0.6325>.
- Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.
- Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. SimulEval: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.
- Dominik Macháček and Peter Polák. 2025. Simultaneous Translation with Offline Speech and LLM Models in CUNI Submission to IWSLT 2025. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 389–398. Association for Computational Linguistics. DOI: [10.18653/v1/2025.iwslt-1.41](https://doi.org/10.18653/v1/2025.iwslt-1.41).
- Kayo Matsushita. 2026. [Understanding AI interpreting in context: A comprehension-based evaluation of human vs. machine-generated interpretations in a real-world setting](#). *International Journal of Language, Translation and Intercultural Communication*, 11(1):71–85. Comprehension-based evaluation of human interpreters vs. the KUDO AI Speech Translator.
- Microsoft. 2026. What is speech translation? Azure AI services documentation. <https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-translation>. Accessed 2026-08-02. Documents the ASR→TrueText→MT→TTS pipeline and its per-stage billing.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- OpenAI. 2022. Whisper large-v3 model card. Hugging Face. <https://huggingface.co/openai/whisper-large-v3>. Accessed 10 September 2026.
- OpenAI. 2026. Model documentation: gpt-realtime-translate. <https://developers.openai.com/api/docs/models/gpt-realtime-translate>. Accessed 2026-08-02. GA May 2026; \$0.034 per audio minute.
- Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. StreamAtt: Direct streaming speech-to-text translation with attention-based audio history selection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3692–3707, Bangkok, Thailand. Association for Computational Linguistics. ArXiv:2406.06097; introduces StreamLAAL.
- Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, pages 12–17, Online. Association for Computational Linguistics.
- Sara Papi, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025. How “real” is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics*, 13:281–313.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.
- 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 (ICML)*, 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 (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, et al. 2023a. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Seamless Communication, Loïc Barrault, Yu-An Chung, et al. 2023b. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*. Introduces SeamlessStreaming and BLASER-2.0.

Kilian G. Seeber and Dongpeng Pan. 2026. Sounding Out Remote Interpreting: The Impact of Sound Quality on Workload and Physiological Arousal. *Perspectives*. In press.

World Health Organization. 2025. Report on WORDLY AI interpretation. Technical report, World Health Organization. Assessment of the Wordly AI interpretation system on 2024 World Health Assembly speeches in the six official languages, conducted by WHO language services staff with the assistance of University of Geneva FTI interpretation teachers; reported lag of over 32 s, with only 1 of 90 interpretations reaching the 75% passing grade. Available at <https://www.atanet.org/wp-content/uploads/2025/06/Report-on-WORDLY-AI-Interpretation.pdf>.

A Reproducibility

Open-arm models and pins. ASR openai/whisper-large-v3; MT facebook/nllb-200-distilled-1.3B; TTS facebook/mms-tts-{fra,deu,spa,rus}, with Italian synthesis via the Piper voice it_IT-paola-medium because facebook/mms-tts-ita was withdrawn from the HuggingFace Hub during the study window (verified 2026-08-03 against the canonical repo, mirrors, and CDN). SeamlessStreaming ran the uniform SeamlessStreamingS2STAgent with vocoder_v2 in all five languages; SeamlessExpressive/PRETSSEL was not used (no Russian support), and every manifest records the effective

agent, asserted uniform before pooling. Software pins: torch==2.2.2+cu121, transformers==4.49.0, silero-vad==6.2.1, simuleval==1.1.4, faster-whisper (large-v3) for all transcription. Open-Cascade ran on NVIDIA Titan RTX hardware with fp16 inference and real-time-paced input. The evaluated natural SeamlessStreaming rerun used a rented A100, as described in Section 4.

Operating-point sweeps (dev speech ARI₁→fr). Cascade, 2026-08-04: {silence 300/500 ms} × {cap 8/12 s} → 500 ms/8 s selected (median per-commit lag 3.2/3.3/2.7 s and max 10.3/6.4/4.4 s for baseline/sil300/cap8). Azure segmentation: semantic default vs. Speech_Segmentation SilenceTimeoutMs=300 (median source segment 6.9→3.8 s, translated words 1657→1683, audibly degraded delivery); default retained, the alternative sweep archived as a latency–integrity trade-off.

Commercial access. gpt-realtime-translate accessed 2026-08-03; Azure Speech Translation in eastus; Speech SDK Translation Recognizer; voices de-DE-KatjaNeural, es-ES-ElviraNeural, fr-FR-DeniseNeural, it-IT-ElsaNeural, ru-RU-SvetlanaNeural, accessed 2026-08-03–04. All commercial runs in a single dated window (August 3–4, 2026). Costs: \$7.25 (gpt-realtime-translate, list price \$0.034/min); \$0 billed (free-tier credits; list-price equivalent ≈\$18–36 for both sweeps) (Azure Speech Translation); open arms ≈9 GPU-h and ≈30 GPU-h of cluster GPU time. Code, run manifests, processed alignment tables, and analysis scripts will be released on publication; rendition audio and ratings remain under the original consent terms and are available through the source experiment’s controlled access route. Released commercial-service outputs carry the usage note “commercial-service outputs: not for training competing models.”

Inclusion and validation. Master inclusion table: 74 of 78 recorded participants included; PF40 (no usable audio), PF30 (dropped in the original quality evaluation), PX19 (two trials), PX20 (withdrew) excluded (Section 3). Alignment hand-validation: complete side-by-side walkthroughs of three renditions per language reviewed before scoring; the automatic alignment matches individual source and target segments one-to-one, so a target segment expressing content from multiple source segments does not automatically credit every source segment. Coverage counts a canonical source segment as matched when a retained

one-to-one alignment links it to a target sentence at similarity ≥ 0.25 under the monotonic program; the four speeches contain 460 canonical segments (108–121 per speech). Latency quantization floors: gpt-realtime-translate ~ 200 ms audio deltas; Azure Speech Translation segment-finalized synthesis; Open-Cascade VAD-commit units (500 ms/8 s); SeamlessStreaming 320 ms input blocks.

Interpreting the SONAR coefficient on the rating scale. We calculated expected content ratings from the fitted cumulative-logit model by weighting the seven response categories by their predicted probabilities. We compared SONAR at its pooled mean with one SD above that mean, retained the language fixed effects, and integrated over the estimated Gaussian random-intercept distributions for rater, interpreter and excerpt. Averaging the language-specific changes in proportion to their numbers of ratings gave 0.23 points (0.22–0.27 across languages). The reported range is variation across languages, not a confidence interval.