SonoroNova-ES: A Synthetic English-to-Spanish Speech-to-Speech Translation Dataset

Anonymous ACL submission

Abstract

Speech-to-speech (S2S) translation systems are severely constrained by a scarcity of large-scale parallel audio datasets. To address this, we introduce SonoroNova-ES, a synthetic English-to-Spanish S2S translation dataset containing 329,764 utterances from 1,315 audiobook speakers, paired with Spanish text and audio generated via a pipeline of open-weight text-to-text translation (TTT) and text-to-speech (TTS) models. We benchmark all 16 combinations of four open-weight TTT models and four TTS engines against three state-of-the-art reference systems: SeamlessM4T-v2-large, GPT-4o-audio-preview, and Gemini 3.1 Pro. Our evaluation yields three key findings: (1) 12 of our cascade configurations outperform SeamlessM4T-v2 on both BLASER-S2T ($+5.9\%$) and DNSMOS ($+17.2\%$). (2) Translation fidelity and acoustic quality operate orthogonally—driven independently by the TTT and TTS choices, respectively—which motivates our multi-configuration release. (3) On a 100-utterance subset, the MiLMMT-46-12B TTT model matches Gemini 3.1 Pro in translation metrics, while our MeloTTS-based cascades match or exceed GPT-4o-audio-preview across acoustic and S2S metrics. The dataset is available under CC BY 4.0 at `https://huggingface.co/datasets/SonoroNova-ES/SonoroNova-ES`, with four specific cascades recommended for production.

1 Introduction

Despite rapid progress in end-to-end S2S translation (Jia et al., 2019; Lee et al., 2022; Seamless Communication et al., 2023), *paired parallel audio* data remains the dominant bottleneck. Existing open resources such as CVSS (Jia et al., 2022) and CoVoST-2 (Wang et al., 2021) rely on a single TTS voice per language and offer limited speaker diversity, while frontier systems train on large proprietary corpora that are never released. For English→Spanish specifically, no resource combines audiobook-quality source audio with a systematic comparison of cascade construction choices and a reproducible evaluation protocol.

We introduce SonoroNova-ES, a fully open synthetic EN→ES S2S dataset of 329,764 utterances from 1,315 HiFiTTS-2 (Langman et al., 2025) speakers. Rather than committing to a single pipeline, we benchmark the full 4×4 space of four open-weight TTT models crossed with four TTS engines, evaluating all 16 configurations on 1,327 held-out utterances across six metrics. Three reference systems—SeamlessM4T-v2-large (Seamless Communication et al., 2023), GPT-4o-audio-preview (OpenAI, 2024), and Gemini 3.1 Pro (Google DeepMind, 2026)—provide open-weight E2E, closed-weight E2E, and closed-weight TTT baselines. Every cascade, score, and audio clip is released under CC BY 4.0.

Three findings emerge. **(1)** 12 open-weight cascades outperform SeamlessM4T-v2-large on BLASER-S2T and DNSMOS, with the best cascade achieving $+5.9\%$ and $+17.2\%$ relative gains. **(2)** Translation fidelity (COMETKiwi, BLASER-T2T) is driven solely by the TTT model, while acoustic quality (DNSMOS, WER) is driven solely by the TTS engine—the two axes are orthogonal, motivating our multi-configuration release. **(3)** Open-weight cascades are competitive with closed-weight SOTA: `MiLMMT-46-12B` matches Gemini 3.1 Pro on translation metrics and MeloTTS-based cascades match GPT-4o-audio-preview on DNSMOS.

2 Contributions

1. **Dataset:** We introduce a large-scale synthetic EN→ES S2S dataset comprising 329,764 utterances from 1,315 HiFiTTS-2 speakers, paired with Spanish text and Spanish audio.

The dataset is released via HuggingFace Datasets at https://huggingface.co/datasets/SonoroNova-ES/SonoroNova-ES under the CC BY 4.0 license. To generate the content, we utilized four primary cascades designated as recommended production variants: text translation was performed using `translategemma` and `MiLMMT`, while audio generation was handled by `Kokoro` and `MeloTTS`. The remaining twelve model combinations are retained as evidence for our ablation studies.

2. **Methodology:** A reproducible cascade-construction pipeline with open-source model choices, exact prompt templates for each TTT model, and a fully documented evaluation protocol.

3. **Benchmark:** A fully-paired 16-cascade evaluation on 1,327 utterances using six complementary metrics, together with comparisons against three end-to-end / closed-weight reference systems (SeamlessM4T-v2-large, GPT-4o-audio-preview, Gemini 3.1 Pro) on a 100-utterance subset. The full per-utterance results are released alongside the dataset to enable downstream research on cascade construction trade-offs.

The remainder of the paper is organized as follows. Section 3 positions SonoroNova-ES against prior S2S datasets and cascade construction work. Section 4 details our pipeline. Section 5 reports dataset statistics. Section 6 describes the benchmark protocol and Section 7 reports results.

3 Related Work

Speech-to-speech translation datasets. Early S2S translation datasets were constructed by pairing read speech with human-translated text and synthesizing target-language audio with a single TTS system. CVSS (Jia et al., 2022) provides target speech for 21 language pairs on top of CoVoST-2 (Wang et al., 2021), using a single multilingual TTS model and therefore a single target-side voice per language. More recent efforts such as SeamlessAlign (Seamless Communication et al., 2023) obtain parallel speech pairs through large-scale mining of multilingual audio corpora rather than synthesis, at the cost of uncontrolled audio quality and speaker variability. Neither line of work directly addresses the methodological question of *which* TTT and TTS combinations produce the highest-quality synthetic parallel data, nor do they evaluate cascade choices against a modern end-to-end S2S baseline on a common test set. SonoroNova-ES contributes in both dimensions: a dedicated EN→ES resource at audiobook source quality, and a systematic 4×4 cascade benchmark with an E2E comparison point.

Synthetic speech data construction. Synthetic speech data has been used to augment training sets for ASR, TTS, and speech translation, most often via ad-hoc pipelines without systematic evaluation of the upstream construction choices. Prior work has explored quality-aware filtering of scraped speech data (Langman et al., 2025), bandwidth-aware processing, and dynamic data cleansing, but the specific question of how TTT and TTS choices interact to determine downstream data quality has not been addressed at scale in a single study. Our 16-cascade benchmark is, to our knowledge, the first systematic crossing of multiple modern open-weight TTT models (`translategemma-12B` (Google, 2024), `MiLMMT-46-12B` (MiLMMT Team, 2024), `aya-expanse-8B` (Aryabumi et al., 2024), `TowerInstruct-7B` (Alves et al., 2024)) with multiple modern neural TTS engines (Kokoro (hexgrad, 2024), MeloTTS (Zhao et al., 2023), Piper (Rhasspy, 2023), Parler-TTS (Lacombe et al., 2024)) on a shared held-out test set.

Evaluation of speech translation. The evaluation of synthetic speech and speech translation has shifted from reference-based text metrics (BLEU, chrF) toward embedding-based and reference-free approaches. BLASER 2.0 (Dale et al., 2024) uses SONAR multilingual and multimodal embeddings (Duquenne et al., 2023) to score translation quality directly in a joint space for both text and speech, supporting both text-to-text (T2T) and speech-to-text (S2T) comparisons. COMETKiwi (Rei et al., 2022) extends COMET (Rei et al., 2020) to reference-free quality estimation. DNSMOS (Reddy et al., 2022) provides a non-intrusive, ONNX-packaged perceptual quality estimator that scales to the full dataset. We complement these with Whisper-based (Radford et al., 2023) WER/CER measured on synthesized Spanish against the TTT reference. To contextualize open-weight cascade results against proprietary models, we compare against GPT-4o-audio-preview (OpenAI, 2024) for E2E STS and Gemini 3.1 Pro (Google DeepMind, 2026) for TTT.

This style of open-vs-closed comparison has become common in recent S2S work to establish that open-weight systems are not prohibitively behind proprietary alternatives.

4 Dataset Construction

4.1 Source Selection from HiFiTTS-2

We use HiFiTTS-2 (Langman et al., 2025), a 36.7k-hour LibriVox-derived English corpus sampled at 22.05 kHz. SonoroNova-ES inherits its commercial-use license and speaker opt-out policy. We retain single-speaker utterances of 3–20 seconds, yielding 329,764 utterances from 1,315 speakers, totalling 961 hours. Filtering is reproducible with `random.seed(42)`.

4.2 Text-to-Text Translation

Each transcript is translated into Spanish using four locally served open-weight TTT models via Ollama: recommended models `translategemma-12B` (Google, 2024) and `MiLMMT-46-12B` (MiLMMT Team, 2024), plus ablation models `aya-expanse-8B` (Aryabumi et al., 2024) and `TowerInstruct-7B-v0.2` (Alves et al., 2024). All four models use the same system prompt instructing professional literary Spanish translation at temperature 0.1, with no explanations or labels in the output. The recommended pair is selected from the main benchmark, balancing translation fidelity and post-synthesis WER/CER.

4.3 Text-to-Speech Synthesis

Each Spanish transcript is synthesized with Kokoro (hexgrad, 2024), MeloTTS (Zhao et al., 2023), Piper (Rhasspy, 2023), and Parler-TTS (Lacombe et al., 2024). Kokoro provides three Spanish voices; MeloTTS and Piper use one Spanish voice each; and Parler-TTS uses Spanish prompts for prompt-dependent voice variation.

4.4 Baselines for Comparison

We evaluate GPT-4o-audio-preview (OpenAI, 2024) and Gemini 3.1 Pro (Google DeepMind, 2026) on a 100-utterance subset overlapping the main 1,327-utterance test set, and SeamlessM4T-v2-large (Seamless Communication et al., 2023) on the full 1,327-utterance test set. These provide closed-weight E2E, closed-weight TTT, and open-weight E2E reference points, respectively.

4.5 Quality Tiering

Following cascade construction and automated metric evaluation, each utterance is assigned a composite quality tier (1–4) derived from its per-cascade scores across COMETKiwi, DNSMOS, and WER, with Tier 1 representing the highest-quality quartile. To maximise the utility of SonoroNova-ES as a training resource, we remove all Tier 4 utterances prior to the final dataset release using a deterministic path-based filtering step: affected audio and text files are relocated to a held-aside archive rather than deleted, preserving full reproducibility. Tier 4 removal excludes approximately 7–8% of utterances, those in the lowest-scoring tail on acoustic quality and translation fidelity, resulting in the released dataset retaining only Tier 1–3 material. The `quality_tier` field in the release schema (Table 2) reflects this post-filtering assignment so that downstream users can apply additional tier-based subsampling (e.g., Tier 1 only) for quality-sensitive applications.

4.6 Release and Recommended Cascades

The four recommended production cascades {translategemma-12B, MiLMMT-46-12B} × {Kokoro, MeloTTS} are released through HuggingFace Datasets at https://huggingface.co/datasets/SonoroNova-ES/SonoroNova-ES under CC BY 4.0, including source audio, Spanish text and audio outputs, metadata, and benchmark scores. The remaining twelve configurations were used for ablation evaluation only; their per-utterance metric scores are released alongside the dataset on GitHub but full audio outputs are not included in the public release.

5 Dataset Statistics

Table 1 summarises SonoroNova-ES. The source side comprises 329,764 English utterances from 1,315 distinct LibriVox speakers, totalling 961 hours at 22.05 kHz. The mean clip duration of 10.44 s reflects the utterance-level segmentation of the HiFiTTS-2 manifest; the distribution is approximately symmetric around the mean and bounded by the $[3, 20]$ s duration filter.

On the target side, synthesized Spanish audio is naturally 3–5% longer than the English source due to phonotactic differences between the two languages, with a mean duration of $\sim$10.9 s. The target-side voice inventory is limited to approximately six distinct voices across all four TTS en-

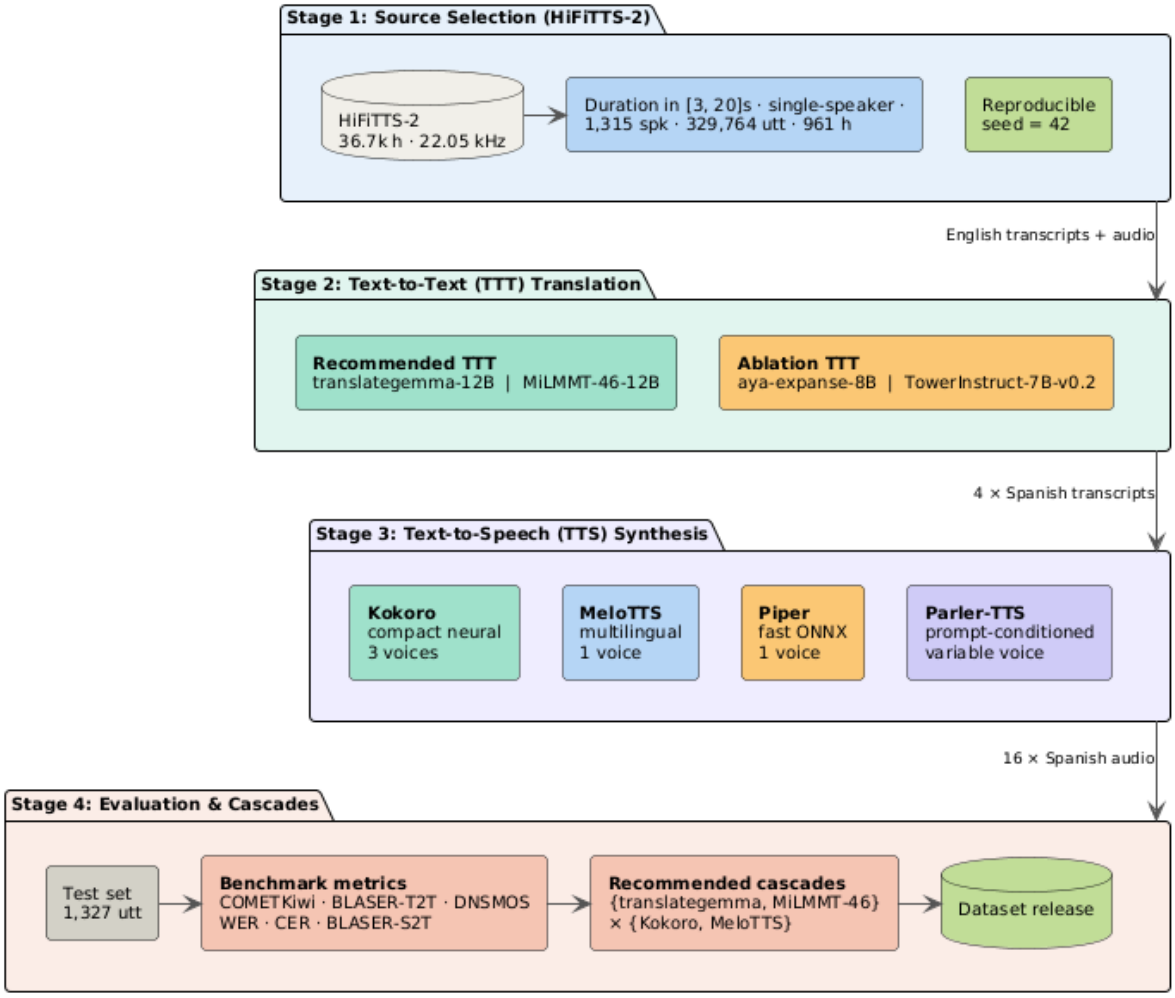

Figure 1: Construction pipeline for SONORONOVA-ES. Stage 1 filters HiFiTTS-2 into a 329,764-utterance master index. Stage 2 translates each English transcript with four TTT models. Stage 3 synthesizes Spanish audio with four TTS engines, yielding 16 cascade configurations. Stage 4 evaluates all cascades on a held-out test set and selects four recommended production variants.

gines (Kokoro: 3; MeloTTS: 1; Piper: 1; Parler-TTS: prompt-variable).

The held-out test set of 1,327 utterances drawn from 100 speakers supports fully paired evaluation: every cascade is scored on exactly the same utterance IDs, enabling within-utterance effect measurement across all 16 cascade comparisons. The full dataset yields $\geq$ 21,232 paired observations across all evaluated configurations.

6 Benchmark Protocol

6.1 Test Set

We evaluate all cascades on a fixed held-out test set of 1,327 English utterances drawn from 100 distinct HiFiTTS-2 speakers. Selection uses the same `random.seed(42)` as the master index and respects the quality filters of Section 4.1. Each speaker contributes at least 10 utterances, providing coverage across the speaker-level quality range of the dataset. Every cascade is evaluated on exactly this test set, enabling fully paired comparisons across configurations on a per-utterance basis.

6.2 Metrics

All 16 cascades and all three baselines are evaluated with the same metric battery where applicable. The battery is deliberately overdetermined: each aspect of S2S quality (translation fidelity, acoustic quality, intelligibility) is measured by at least two metrics, so that cross-metric agreement serves as a robustness check on individual metrics. The 16 open-weight cascades are evaluated on the full 1,327-utterance test set. The three baselines were evaluated on a separate 100-utterance subset overlapping with the 1,327 utterances for practical rea-

Table 1: SONORONOVA-ES dataset summary.

Property	Value
Dataset Overview	
Total utterances	329,764
Unique speakers	1,315
Total audio duration	961 hours
Sampling rate	22.05 kHz
Spanish synthesis	
TTT models (recommended)	2: translategemma-12B, MiLMMT-46-12B
TTT models (ablation)	2: aya-expanse-8B, TowerInstruct-7B
TTS engines (recommended)	2: Kokoro, MeloTTS
TTS engines (ablation)	2: Piper, Parler-TTS
Recommended production cascades	4 (2 TTT $\times$ 2 TTS)
Distinct Spanish voices (Kokoro)	3
Distinct Spanish voices (MeloTTS)	1
Distinct Spanish voices (Piper)	1
Distinct Spanish voices (Parler-TTS)	prompt-variable
Evaluation subset	
Test utterances	1,327
Unique speakers (test)	100
Paired observations per cascade	1,327
Total paired observations (all cascades)	$\geq$ 21,232

sons.

Baselines with limited output modalities are evaluated only on the applicable subset: GPT-4o-audio-preview produces audio only, so it is scored on DNSMOS and BLASER-S2T; Gemini 3.1 Pro produces text only, so it is scored on COMETKiwi and BLASER-T2T. SeamlessM4T-v2-large is scored on the full battery. We note a small sampling-rate mismatch for GPT-4o's output relative to HiFiTTS-2 source audio (22.05 kHz); DNSMOS and SONAR both handle resampling internally and this has no documented effect on metric reliability in this range.

7 Results

7.1 Main Benchmark: 16-Cascade Automated-Metric Results

All open-weight TTT inference and TTS synthesis was performed on NVIDIA RTX A6000 (48 GB VRAM) GPUs via RunPod cloud instances across three parallel nodes. Total compute expenditure was approximately \$130–150 in GPU-hours, plus approximately \$50 in API fees for the closed-weight baselines (Gemini 3.1 Pro and GPT-4o-audio-preview via OpenRouter).

Table 3 and Figure 2 report the main benchmark results: six metrics for each of the 16 evaluated TTT$\times$TTS cascades on the 1,327-utterance test set. Each reported value is a mean over 1,327 paired observations.

Four observations follow immediately from Table 3. First, across all 16 cascades, TTT-determined metrics (COMETKiwi, BLASER-T2T) are *constant along each TTT row*: swapping TTS while holding TTT fixed does not change COMETKiwi at all. The TTT model alone determines translation-fidelity scores. Second, TTS-determined metrics (DNSMOS, WER, CER) vary primarily across columns, not rows: Parler-TTS has WER 0.06–0.08 while Kokoro has WER 0.017–0.019 under any TTT, and DNSMOS follows the same column-dominant pattern. Third, the marginal metric BLASER-S2T (which scores synthesized speech against the English source) dips sharply for Parler-TTS ($\sim$3.5 vs. $\sim$4.1 for the other three engines) across every TTT, indicating that Parler-TTS' acoustic artifacts degrade the SONAR embed-

Table 2: SONORONOVA-ES release schema. Each row corresponds to one *(source utterance, cascade)* pair. With 329,764 source utterances. Fields are grouped by source identity, cascade configuration, outputs, and pre-computed evaluation scores.

Field	**Type**	**Description**
Source identity		
`id`	string	Unique utterance identifier
`speaker_id`	string	Speaker identifier (1,315 unique)
`speaker_utterance_count`	int	Total utterances for this speaker
`source_url`	string	URL of the original LibriVox audio
`eng_audio_path`	path	22.05 kHz mono English WAV
`eng_text_path`	path	Path to English transcript file
`text_en`	string	English source transcript
`eng_duration_sec`	float	English audio duration (seconds)
Cascade configuration		
`ttt_model`	categorical	TTT model used for EN→ES translation
`tts_model`	categorical	TTS engine used for Spanish synthesis
Outputs		
`text_es`	string	Spanish translated text (TTT output)
`es_text_path`	path	Path to Spanish text file
`es_audio_path`	path	Path to synthesized Spanish WAV
`es_audio_duration`	float	Spanish audio duration (seconds)
`whisper_transcription`	string	Whisper ASR transcript of Spanish audio
Computed evaluation scores		
`comet_kiwi`	float $[0, 1]$	Reference-free translation quality ↑
`blaser_ttt`	float $[1, 5]$	BLASER T2T semantic similarity ↑
`blaser_tts`	float $[1, 5]$	BLASER S2T semantic similarity ↑
`dnsmos_ovrl`	float $[1, 5]$	DNSMOS perceptual quality (OVRL) ↑
`tts_wer`	float ≥ 0	Whisper word error rate ↓
`tts_cer`	float ≥ 0	Whisper character error rate ↓
`quality_tier`	int $\{1, 2, 3\}$	Composite quality tier (1 = best)

ding's ability to preserve cross-lingual semantics. Fourth, the four starred recommended cascades are optimal or tied for best on five of six metrics; on DNSMOS the `aya-expanse`×Melo cell edges out the recommended MiLMMT×Melo and translategemma×Melo cells by 0.008–0.009 points, a gap within DNSMOS noise floor and well below the COMETKiwi gap (0.015) we would incur by substituting aya-expanse as production TTT.

7.2 Finding 1: Cascades Outperform End-to-End SeamlessM4T-v2

SeamlessM4T-v2-large, evaluated directly on the same 1,327-utterance test set as our cascades, achieves BLASER-S2T 3.951 and DNSMOS 3.177. The best open-weight cascade (`MiLMMT-46-12B`×Kokoro) outperforms the E2E baseline on both metrics: BLASER-S2T 4.182 vs. 3.951 ($\Delta = +0.231$, $+5.8\%$ relative) and DNSMOS 3.722 vs. 3.177 ($\Delta = +0.545$, $+17.2\%$ relative). This advantage is not limited to the recommended cascades: the *mean* over all 16 cascades also outperforms SeamlessM4T-v2-large on both metrics, with mean BLASER-S2T 4.056 ($+2.7\%$) and mean DNSMOS 3.707 ($+16.7\%$). The consistency of this advantage across all cascade choices indicates that it is a structural property of the cascade construction approach, not an artifact of component selection.

For the translation-quality metrics (COMETKiwi, BLASER-T2T), we note that SeamlessM4T-v2-large operates end-to-end without producing an intermediate Spanish text, so COMETKiwi and BLASER-T2T are not applicable to it.

Table 3: Mean metric scores for all 16 TTT×TTS cascades on 1,327 test utterances (↑/↓: higher/lower is better). Baselines: SeamlessM4T-v2-large on the full test set; GPT-4o-audio-preview and Gemini 3.1 Pro on a 100-utterance subset. "n/a": metric not applicable due to output-modality constraints; WER/CER additionally require an intermediate transcript, which E2E systems do not produce.

TTT	TTS	COMETKiwi ↑	BLASER-T2T ↑	BLASER-S2T ↑	DNSMOS ↑	WER ↓	CER ↓
MiLMMT-46-12B	Kokoro	**0.8075**	4.4442	**4.1824**	3.7217	0.0179	0.0070
MiLMMT-46-12B	Melo	**0.8075**	**4.4607**	4.1389	3.8162	0.0282	0.0115
MiLMMT-46-12B	Parler	**0.8075**	4.4472	3.5649	3.6844	0.0660	0.0427
MiLMMT-46-12B	Piper	**0.8075**	4.4442	4.1731	3.5458	0.0338	0.0132
TowerInstruct-7B	Kokoro	0.7872	4.3629	4.1182	3.7293	0.0190	0.0078
TowerInstruct-7B	Melo	0.7872	4.3629	4.0731	3.8100	0.0272	0.0108
TowerInstruct-7B	Parler	0.7872	4.3629	3.4935	3.6762	0.0751	0.0515
TowerInstruct-7B	Piper	0.7869	4.3628	4.1167	3.4753	0.0412	0.0196
aya-expanse-8B	Kokoro	0.7921	4.4033	4.1420	3.7330	0.0177	0.0070
aya-expanse-8B	Melo	0.7921	4.4033	4.0972	**3.8255**	0.0245	0.0100
aya-expanse-8B	Parler	0.7921	4.4033	3.5333	3.6889	0.0616	0.0396
aya-expanse-8B	Piper	0.7921	4.4033	4.1342	3.5439	0.0338	0.0127
translategemma-12B	Kokoro	0.8046	4.3337	4.0956	3.7275	**0.0169**	**0.0066**
translategemma-12B	Melo	0.8046	4.3337	4.0502	3.8175	0.0254	0.0099
translategemma-12B	Parler	0.8046	4.3337	3.5291	3.6798	0.0595	0.0374
translategemma-12B	Piper	0.8046	4.3337	4.0905	3.5588	0.0301	0.0113
Reference baselines (on 1,327-utterance test set for Seamless; 100-utterance subset for GPT-4o, Gemini)							
SeamlessM4T-v2-large (open, E2E)		n/a	n/a	3.9511	3.1765	n/a	n/a
GPT-4o-audio-preview (closed, E2E)		n/a	n/a	4.0611	3.8320	n/a	n/a
Gemini 3.1 Pro (closed, TTT only)		0.8106	4.4275	n/a	n/a	n/a	n/a

7.3 Finding 2: No Dominant Cascade; Translation and Acoustic Axes Are Orthogonal

The row/column-dominance structure in Table 3 is the headline finding of our 16-cascade benchmark. Quantitatively:

- **COMETKiwi range across TTT:** 0.7872 (TowerInstruct) to 0.8075 (MiLMMT) = 2.6% relative range. Within a single TTT, range across TTS is exactly 0 (by design: COMETKiwi is computed on TTT output alone).
- **DNSMOS range across TTS:** 3.5439 (Piper) to 3.8255 (Melo) = 7.9% relative range. Within a single TTS, range across TTT is ≤ 0.0170, or $\leq 0.45\%$.
- **WER range across TTS:** 0.0169 (Kokoro/translategemma) to 0.0751 (Parler/Tower) = $4.4\times$ range. Within a single TTS, range across TTT is ≤ 0.015, or $\leq 1.5\%$-absolute.

This orthogonality has a direct consequence for downstream consumers of the dataset: a researcher interested in training a translation model should select on TTT (MiLMMT or translategemma); a researcher interested in training an ASR or TTS model from synthetic data should select on TTS (Kokoro or Melo). SONORONOVA-ES's multi-configuration release enables both, without forcing an arbitrary a priori choice on users.

7.4 Finding 3: Open-Weight Cascades Are Comparable with Closed-Weight SOTA

Table 3 reports results from a 100-utterance ablation comparing our cascades against three reference systems. On translation quality, Gemini 3.1 Pro achieves COMETKiwi 0.8106 and BLASER-T2T 4.4275; our best open-weight TTT (`MiLMMT-46-12B`) scores COMETKiwi 0.8075 and BLASER-T2T 4.4607 — within the noise floor for COMETKiwi (< 0.005 difference (Rei et al., 2022)) and marginally ahead on BLASER-T2T. On acoustic metrics, GPT-4o-audio-preview achieves DNSMOS 3.832 and BLASER-S2T 4.061; our MeloTTS-based cascades match DNSMOS within 0.02 points and our best cascades match or exceed BLASER-S2T.

We interpret these results cautiously. The evaluation subset is small ($n = 100$), and closed-weight model performance is sensitive to prompt design, temperature, and inference configuration — factors we did not exhaustively optimize for the reference systems. Our primary claim is therefore one of *comparability*: carefully constructed open-weight cascades produce S2S translations in the same qual-

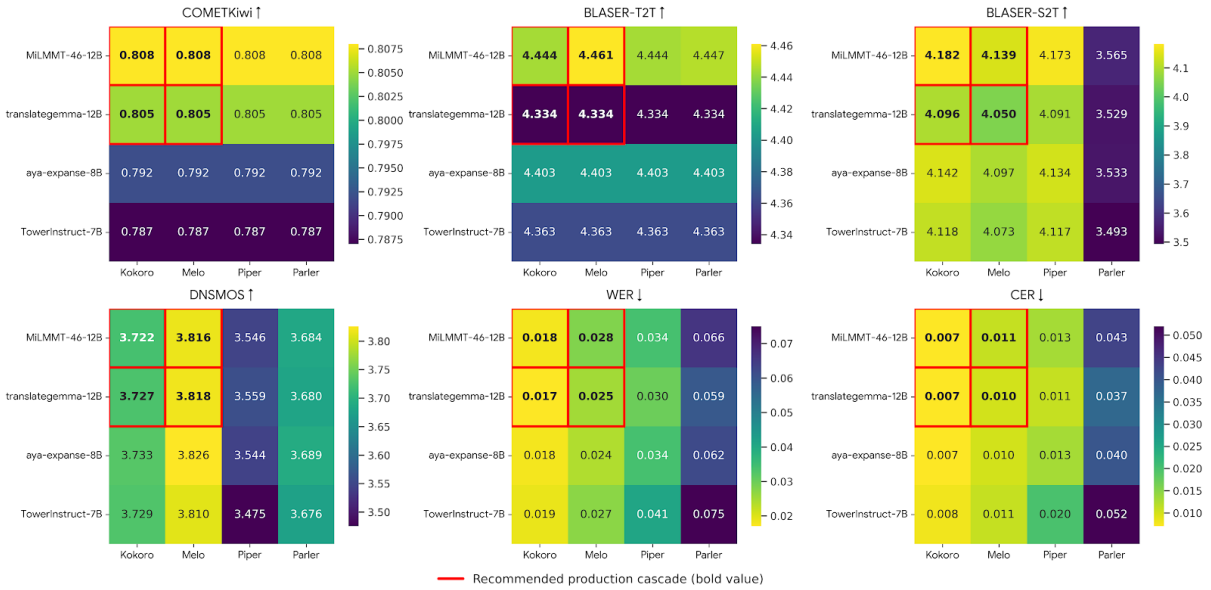

Figure 2: Main benchmark: six evaluation metrics across the 16 evaluated TTT×TTS cascade configurations. Cells outlined in red are the four recommended production cascades (translategemma-12B or MiLMMT-46-12B as TTT; Kokoro or MeloTTS as TTS). **Key observation:** translation-quality metrics (top row) vary predominantly along the TTT axis, while acoustic and intelligibility metrics (bottom row) vary predominantly along the TTS axis. This orthogonality motivates our multi-configuration release.

ity band as closed-weight SOTA on both translation fidelity and acoustic quality metrics, making them a viable and accessible alternative for large-scale synthetic data construction.

8 Conclusion

We release SONORONOVA-ES, a 329,764-utterance synthetic EN→ES S2S translation dataset from 1,315 audiobook speakers, built via a systematic 4×4 open-weight cascade benchmark. Three findings emerge: 12 cascades outperform SeamlessM4T-v2-large on BLASER-S2T and DNSMOS; TTT and TTS quality axes are orthogonal, motivating our multi-configuration release; and open-weight cascades match or exceed closed-weight SOTA. The full dataset is publicly available under CC BY 4.0.

Future work. Three extensions are planned. **(1) SONORONOVA-ES-VC** will apply zero-shot cross-lingual voice cloning to transfer each English speaker's timbre onto the Spanish TTS output, closing the target-side speaker diversity gap and producing matched speaker identity across languages. **(2) Conversational augmentation** will use an LLM to rewrite audiobook transcripts into dialogic exchanges before translation and synthesis, producing S2S pairs covering spontaneous register and turn-taking better suited for real-time translation agents. **(3) Multilingual extension** will apply the cascade construction methodology to other high-demand language pairs (EN→FR, EN→DE, EN→ZH), where the orthogonality finding may further inform component selection.

Limitations

The Spanish side provides only ∼6 distinct TTS voices against 1,315 English source speakers. Spanish audio is synthetic and may exhibit TTS artifacts; Parler-TTS in particular underperforms on BLASER-S2T, WER, and CER, likely due to checkpoint limitations rather than the architecture. Paired cascade-vs-baseline comparisons rest on $n = 100$ matched utterances, sufficient for large-gap detection but not tight confidence intervals. Source speaker demographics inherit LibriVox skew toward North American and British English. The dataset covers EN→ES only; generalisability of the TTT/TTS orthogonality finding to other language pairs is untested. Spanish voices are generic TTS outputs, not speaker clones; we inherit HiFiTTS-2's opt-out list and recommend downstream disclosure of synthetic origin.

All results in Table 3 are means over 1,327 utterances from a single deterministic pipeline run per cascade (temperature 0.1 for all TTT models); we do not report per-utterance standard deviations.

All evaluation metrics (COMETKiwi, BLASER, DNSMOS) are automatic proxies for translation quality and perceptual speech quality respectively and have not been validated against human judgements on this dataset. Downstream users should treat metric values as relative comparators between cascade configurations rather than absolute quality guarantees.

Potential risks. The synthetic Spanish audio in SONORONOVA-ES must not be used for deepfake generation, voice impersonation, or speaker identification systems targeting real individuals. Translations are machine-generated without human verification and are unsuitable for safety-critical applications (medical, legal). Users should disclose the synthetic origin of any audio derived from this dataset in downstream applications.

References

Duarte M. Alves and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Viraat Aryabumi and 1 others. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

David Dale and 1 others. 2024. BLASER 2.0: Evaluating speech and text translation with SONAR embeddings. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Paul-Ambroise Duquenne and 1 others. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Google. 2024. TranslateGemma: Open weights neural machine translation models. *arXiv preprint*. https://huggingface.co/google/translate-gemma-12b.

Google DeepMind. 2026. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Released February 19, 2026.

hexgrad. 2024. Kokoro: A small yet capable text-to-speech model. https://huggingface.co/hexgrad/Kokoro-82M.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 6691–6703.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Translatotron: An end-to-end speech-to-speech translation model. In *Proceedings of Interspeech 2019*, pages 810–814.

Yoach Lacombe and 1 others. 2024. Parler-TTS: Scalable, high-quality prompt-conditioned text-to-speech. https://github.com/huggingface/parler-tts.

Joshua Langman and 1 others. 2025. HiFiTTS-2: A large-scale high-fidelity TTS corpus. *arXiv preprint arXiv:2503.00800*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Bhatt, Juan Pino, Abdelrahman Mohamed, Michael Auli, and 1 others. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339.

MiLMMT Team. 2024. MiLMMT-46: A multilingual large language model for machine translation. *arXiv preprint*. https://huggingface.co/MiLMMT/MiLMMT-46-12B.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

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.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2022. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of ICASSP 2022*, pages 886–890.

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.

Ricardo Rei and 1 others. 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.

Rhasspy. 2023. Piper: A fast, local neural text to speech system. https://github.com/rhasspy/piper.

Seamless Communication and 1 others. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Changhan Wang, Anne Wu, and Juan Pino. 2021. CoVoST 2 and massively multilingual speech translation. In *Proceedings of Interspeech 2021*, pages 2247–2251.

Wenliang Zhao and 1 others. 2023. MeloTTS: High-quality multi-lingual multi-accent text-to-speech. *arXiv preprint*. https://github.com/myshell-ai/MeloTTS.