

NaijaS2ST: A Multi-Accent Benchmark for Speech-to-Speech Translation in Low-Resource Nigerian Languages

Marie Maltais^{1,2}, Yejin Jeon^{1,2}, Min Ma³, Shamsuddeen Hassan Muhammad^{4,5},
Idris Abdulmumin^{4,6}, Maryam Ibrahim Mukhtar⁴, Daud Abolade⁷, Joel Okepefi⁸,
Johnson Sewedo⁸, David Ifeoluwa Adelani^{1,2,9}

¹Mila - Quebec AI Institute, ²McGill University, Canada ³Google DeepMind, ⁴Hausa NLP,
⁵Imperial College, United Kingdom, ⁶University of Pretoria, South Africa, ⁷Masakhane NLP,
⁸Naijá Wikipedia Community, and ⁹Canada CIFAR AI Chair
{marie.maltais, yejin.jeon, david.adelani}@mila.quebec

Abstract

Speech translation for low-resource languages is constrained by the scarcity of high-quality parallel speech data, which is particularly pronounced in African linguistic contexts. To address this, we introduce **NaijaS2ST**, a parallel speech translation dataset covering Igbo, Hausa, Yorùbá, and Nigerian Pidgin paired with English. The dataset contains more than 40 hours of speech per language and includes several speaker and accent variations to reflect realistic settings. Using NaijaS2ST, we benchmark cascaded, end-to-end (E2E), and AudioLLM-based approaches across bidirectional translation directions (XX↔English). Our results show that few-shot audio LLMs are more effective than cascaded and E2E methods finetuned for speech-to-text translation. For speech-to-speech translation, however, cascaded and audio LLM paradigms perform comparatively, highlighting the need for stronger task-specific models. NaijaS2ST offers a strong foundation towards advancing research on low-resource multilingual speech translation. We open source the dataset here: [NaijaS2ST Huggingface Dataset](#); and the code here: [NaijaS2ST Code](#).

1 Introduction

Translation technologies play a central role in enabling equitable access to information and facilitating communication regardless of language boundaries (Naveen and Trojovský, 2024). In recent years, advances in neural architectures have led to substantial improvements in both speech-to-text translation (S2TT) and speech-to-speech translation (S2ST). More importantly, these improvements have not been driven by architectural enhancements alone, but have historically depended on the availability of large-scale parallel corpora.

These advancements, however, have been disproportionately concentrated on a small set of high-

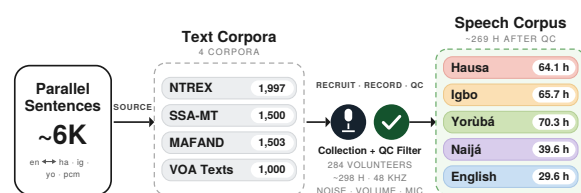


Figure 1: **Construction of the NaijaS2ST corpus.** 6K English-aligned sentences are drawn from four parallel text corpora and read by 284 participants, resulting in ~269 hours of speech after quality control (QC).

resource languages. As a result, what started as a means to reduce language barriers and enable broader access to information and communication has not fully realized its goals, as the benefits of modern translation systems remain inaccessible to a large portion of the world’s population. This disparity is especially evident for African languages, which account for a significant portion of global linguistic diversity. In fact, among the roughly 2,000 languages in the widely used Flores-101 (Goyal et al., 2022) dataset, only 0.85% are African. The more recent FLORES-200 added Hausa, Igbo and Yorùbá, among other African languages, though not Nigerian Pidgin (Costa-jussà et al., 2024).

Recent efforts have led to the development of large-scale speech datasets for Nigerian languages (Meyer et al., 2022; Ògúnremí et al., 2024; Emezue et al., 2025; Adebara et al., 2026). However, these resources primarily target Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) tasks. While there are some prior works for text-based translation (Adelani et al., 2022b; Costa-jussà et al., 2024), it does not address the speech modality, and general-domain datasets capturing African–English accents remain particularly scarce (Olatunji et al., 2023). Moreover, speech-based translation, especially in bidirectional settings, remains significantly underexplored,

as many existing evaluations are limited to the “XX–English” direction in a *single* accent (Wang et al., 2021b; Conneau et al., 2022a).

This lack of data coverage reflects a deeper issue: under-representation of African languages constitutes a fundamental bottleneck for progress in speech translation research. Modern S2TT and S2ST systems, whether based on cascaded pipelines or end-to-end (E2E) architectures, depend on large-scale parallel speech data. In the absence of such datasets, it becomes difficult to (i) perform standardized and reproducible benchmarking, (ii) develop robust and generalizable models, and (iii) conduct meaningful comparisons across different modeling approaches.

We address these challenges by introducing the **NaijaS2ST** speech dataset of actual spoken recordings across the most populous Nigerian languages of Hausa, Igbo, Nigerian Pidgin (Naijá¹), and Yorùbá. This dataset is designed to support both S2TT and S2ST, and includes parallel speech aligned with English. The speech data were collected from 284 participants, who were asked to read 6,000 sentences sourced from various news outlets. Figure 1 illustrates the overall data collection process.

In addition to dataset construction, a systematic benchmarking study of various speech translation models is conducted. Specifically, we evaluate both cascaded approaches and E2E models across different translation directions (English↔African languages). This comprehensive evaluation framework enables a direct comparison of modeling paradigms under low-resource conditions and provides insights into their relative strengths and limitations. We hope that this work will serve as a foundation for future research in speech translation for African languages and contribute to more inclusive multilingual technologies.

2 Related Work

The evolution of speech translation research has transitioned from cascaded architectures toward unified, E2E multimodal paradigms. Yet, such architectural progress has simultaneously exposed structural bottlenecks regarding data equity, accent robustness, and cross-lingual generalizability in low-resource settings (Sarim et al., 2025).

In particular, the rapid advancement of speech translation architectures is fundamentally rooted

in the availability of large-scale datasets. As such, the foundation of modern systems relies heavily on self-supervised learning (SSL) frameworks, such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), which utilize massive unannotated corpora like VoxPopuli (Wang et al., 2021a) and Common Voice (Ardila et al., 2020) for cross-lingual acoustic representations. For parallel translation tasks, monumental benchmarks such as CoVoST 2 (Wang et al., 2021b) and FLEURS (Conneau et al., 2022b) established standard evaluation protocols, while the CVSS corpus (Jia et al., 2022) and the mined SpeechMatrix (Duquenne et al., 2023) expanded bidirectional S2ST capabilities. However, current corpora are predominantly focused on high-resource Indo-European languages. Recent speech datasets for Nigerian languages (Meyer et al., 2022; Ògúnṛémí et al., 2024; Emezue et al., 2025), are also primarily restricted to ASR or TTS.

Moreover, as the S2TT research paradigm has increasingly integrated Large Language Models (LLMs), this has further amplified resource requirements (Rubenstein et al., 2023; qian Dong et al., 2024; Zhang et al., 2024; Barrault et al., 2025; Alastruey et al., 2026). Concurrently, direct speech-to-speech translation has shifted toward textless modeling. While early systems mapped continuous spectrograms (Jia et al., 2019), later work introduced discrete acoustic units (Lee et al., 2022a,b) that enabled models to bypass intermediate text generation and support applications in unwritten languages (Zhang et al., 2021; Chen et al., 2023). Despite this, because these approaches still require substantial parallel speech data, strategies such as cross-lingual pseudo-labeling (Dong et al., 2022; Popuri et al., 2022) and monolingual data utilization (Nachmani et al., 2024) have been utilized to reduce this dependency. Regardless, their LLM reliance results in hallucinations when processing noisy and accented inputs (Sarim et al., 2025).

Despite significant architectural advances, progress in speech translation remains limited by the absence of diverse, multi-accent, and bidirectional parallel speech corpora for low-resource languages. Moreover, evaluating whether discrete-unit or LLM-based E2E models truly outperform cascaded pipelines requires rigorous benchmarking on authentic low-resource data, rather than high-resource standardized proxies. Such work has been presented in high-resource language scenarios, as in Bentivogli et al. (2021), but still lack evaluation in low-resource African languages. To address

¹Naijá is a popular nickname for Nigeria, and also an alternative name for Nigerian-Pidgin (Naijá).

Language	Language Family	No. of Speakers	Existing Data (Exist.)	Newly Added (Added)	No. Exist.	No. Added
Hausa (ha)	Afro-Asiatic/Chadic	94M	NTREX, SSA-MT & MAFAND	VOA Texts	5,000	1,000
Igbo (ig)	Niger-Congo/Volta-Niger	34M	NTREX & SSA-MT	VOA Texts & MAFAND	3,497	2,503
Naijá (pcm)	English-based Creole	121M	–	VOA Texts, MAFAND & SSA-MT	–	4,000
Yorùbá (yo)	Niger-Congo/Volta-Niger	54M	NTREX, SSA-MT & MAFAND	VOA Texts	5,000	1,000

Table 1: **NaijaS2ST languages and data source**, including language families, and number of speakers. Existing dataset and their number of instances include: NTREX (1,997) (Federmann et al., 2022), SSA-MT (1,500) (Li et al., 2025), and MAFAND (1503 out of 1925 parallel ha-yo sentences) (Adelani et al., 2022b).

this, we introduce **NaijaS2ST**, a multi-accent parallel speech translation benchmark for Nigerian languages, providing the data foundation necessary for the evaluation of speech translation paradigms under realistic low-resource conditions.

3 NaijaS2ST Dataset

NaijaS2ST is a parallel speech–text benchmark for training and evaluating *speech-to-text* translation and *speech-to-speech* translation models across the five most populous Nigerian languages and their accents. The languages include English (British & Nigerian accent), Hausa, Igbo, Nigerian Pidgin (Naijá), and Yorùbá, each with at least 30 million native speakers. Unlike the other Nigerian languages, Naijá is an English-based Creole language that combines *simplification* for broader accessibility (e.g. less educated people) with lexical influence from English and some local languages. This simplified and mixed linguistic structure poses unique challenges for automatic text and speech translation systems. Further information on language characteristics are described in (§3.1)

Overall, **NaijaS2ST** enables multi-way S2ST and S2TT for a population of over 300 million native speakers. Details on the languages, their speaker population, and the original data source for data curation are shown in Table 1.

3.1 Language Characteristics

All languages in NaijaS2ST make use of the Latin script and strictly follow the same Subject-Verb-Object word order as in English. The language specific characteristics are provided below.

Hausa (ha) is spoken by more than 94M people across West and Central Africa and is native to Nigeria and the Niger Republic. It employs a Latin-based orthography consisting of 44 letters, including special characters such as ɓ, ɗ, and ƙ. Hausa is a tonal language with high and low tones, and exhibits agglutinative morphological structure. It belongs to the Afro-Asiatic language family, which

includes Arabic and Hebrew. Although multiple dialects exist, this work focuses on the dominant Kano dialect spoken in Nigeria, with all recordings collected in the city of Kano.

Igbo (ig) is spoken by more than 34M people in the South-Eastern part of Nigeria. Igbo makes use of 34 Latin-based letters, including diacritics (grave (“`”), acute (“’”) accents, and underdots on “ĩ”, “õ”, and “ụ”). However, in modern times, upper diacritics are often ignored. Igbo is both tonal and agglutinative, and from the Naijá-Congo/Volta-Niger language family.

Nigerian-Pidgin (Naijá) (pcm) is an English-based creole language native to Nigeria and part of the broader West African Pidgin English (WAPE) continuum. Naijá is spoken by over 121M people and is among the top 20 most spoken languages in the world. Despite this, it remains poorly represented in existing corpora. Owing to its similarity to WAPE, it is often broadly labeled as WAPE in prior datasets, leading to inconsistencies in representation. In this paper, the orthographic conventions used in Wikipedia are adopted, following the recommendations in Adelani et al. (2025).

Yorùbá (yo) is native to South-Western Nigeria, Benin and Togo, and spoken by more than 54M people. Yorùbá has 25 Latin letters, excluding the Latin characters (c, q, v, x and z), and instead includes additional characters (ẹ, gb, ẹ, ọ). Yorùbá is a highly isolating language and belongs to the Naijá-Congo/Volta-Niger like Igbo. Yorùbá is a tonal language with three tones: low (“`”), middle (“–”, optional) and high (“’”). The tonal marks and underdots are very important in pronunciation of words and generating correct sounds; their absence often lead to worse results on downstream speech tasks such as TTS (Ògúnremí et al., 2024).

3.2 Text Data Collection

Statistics of the text data are shown in Table 1. To obtain parallel sentences across language pairs, we use three existing resources: NTREX (Federmann

Language	Train split (# hours)		Dev-Test split		Gender ratio Female:Male
	# Recorded	After QC	# Dev	# Test	
English (en)	30.09	24.96	2.33	2.32	F:19, M:21
Hausa (ha)	58.34	55.06	3.89	5.11	F:47, M:25
Igbo (ig)	65.57	56.89	4.43	4.38	F:47, M:25
Naijá (pcm)	41.04	31.30	3.67	4.61	F:13, M:20
Yorùbá (yo)	63.39	61.38	4.38	4.49	F:46, M:26

Table 2: **NaijaS2ST Speech information** before and after Quality Control (QC).

et al., 2022) for Hausa–Igbo, SSA-MT (Li et al., 2025) for Igbo–Yorùbá, and MAFAND (Adelani et al., 2022b) for Hausa–Yorùbá.

To mitigate data contamination, an additional set of 1,000 sentences is collected for English from the VOA website. These sentences are not parallel to any of the training data and are equally balanced across Nigerian and United Kingdom contexts. This is to enable assessment of S2ST models in handling named entity pronunciation across diverse geographical settings. Moreover, 1,000 evaluation sentences are collected each for Hausa and Yorùbá. For Igbo and Naijá, we supplemented the evaluation data by translating the missing portions (i.e., MAFAND for Igbo and the full set for Naijá). In total, 5,000 sentences are collected for training and 1,000 are reserved for evaluation.

3.3 Speech Data Collection

Speaker Recruitment We recruited language coordinators (LCs) who work in academic environments for each of the Nigerian languages. Each LC further recruits speakers within the same city, mostly from their home university, community, or closed family and friends. For Hausa, Igbo and Yorùbá, we recruited 72 speakers mostly from Kano, Anambra/Imo and Lagos/Ogun states respectively to record the text data collected via the procedure detailed in §3.2. Each volunteer recorded 250 utterances, covering 6,000 sentences in total, with every sentence recorded three times.² Finally, for English (Naijá accent), speakers from two different regions in Nigeria were recruited across all data sources except NTREX, with half from the North (Kano) and half from the South (primarily Lagos and environs), involving 32 speakers in total. For NTREX, spoken English data was also collected in the same two accents with eight more speakers.³ The train, development and test split speakers are completely distinct.

²Each speaker was paid \$15 per 250 utterances recorded.

³Due to budget constraints, we did not prioritize obtaining both accents for all NTREX sentences, as much of the content is more Western-oriented.

Recording Tool Since annotators were geographically dispersed, the Telegram mobile app was used to record utterances for practical reasons. All coordination and project management was conducted via WhatsApp as it is a widely used communication platform in Nigeria. Audio files are recorded at a sample rate of 48 kHz and a signal-to-noise ratio of at least 30 dB. Table 2 provides the details on the number of recordings, number of hours and gender distribution. See Appendix B for details.

Quality Control Quality control (QC) was conducted systematically for each reader and language for the dev and test sets. Specifically, 3 to 5 utterances per reader were sampled to assess both speech quality (e.g., naturalness, repetitions, loudness) and recording conditions (e.g., background noise, microphone quality). The most commonly identified problems included excessive background noise, poor microphone quality, and low volume. Given this, utterances from readers who consistently exhibited such issues were discarded and re-recorded by new participants. All problematic recordings in the development and test sets were fully re-recorded. For budgetary reasons, re-recording was not conducted for the train set. Instead, priority was given to ensuring completely clean dev and test set curation. Table 2 summarizes the speech dataset statistics before and after QC.

4 Experimental Settings

4.1 Models

Cascaded Methods Cascaded S2TT and S2ST systems involve two steps: (1) *ASR transcription* and (2) *Translation* into the target language. In S2ST, there is a *third stage of “audio synthesis,”* where translated text is subsequently converted into speech via a TTS model. For ASR, we employ the Omnilingual-ASR 1B LLM model (Team et al., 2025). For translation, we evaluate two systems: NLLB-200 3.3B (Costa-jussà et al., 2024) trained exclusively for MT, and Tiny-Aya-Global 3B (Salamanca et al., 2026), which supports three of our Nigerian languages, but does not natively support Naijá. We make use of 5-examples from the DEV set for few-shot prompting of TinyAya, Gemma 3n-E2B-it (Gemma 3) and Gemma 4-E4B-it (Gemma 4) (see Appendix D for the prompt). All S2ST experiments use Gemini 2.5 Flash TTS.

End-to-End Methods Unlike cascading pipelines, E2E S2TT and S2ST jointly learn direct mappings from speech inputs to either text or

			XX → Eng					Eng → XX				
Method	Model		Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
CASCADED (ASR + MT)	Omnilingual-ASR	+ NLLB	54.1	42.9	N/A	50.6	–	47.6	52.6	N/A	58.0	–
		+ Tiny Aya	52.5	40.7	54.2	35.5	45.7	39.0	40.0	15.7	26.3	30.3
		+ Gemma 4 Zero-Shot	44.6	30.2	<u>53.9</u>	37.2	41.5	40.9	23.6	49.5	28.3	35.6
		+ Gemma 4 Few-Shot	63.3	<u>50.6</u>	34.1	56.7	<u>51.2</u>	42.1	26.8	<u>52.8</u>	28.6	37.6
END-TO-END	SeamlessM4T	Zero-Shot	14.6	20.6	44.6	<u>57.0</u>	34.2	N/A	53.1	N/A	55.4	-
		Mono FT	<u>54.9</u>	52.4	50.7	60.3	54.6	12.1	<u>54.8</u>	61.4	68.6	<u>49.2</u>
		Multi FT	46.7	47.4	56.0	54.3	51.1	53.4	64.6	42.6	<u>68.5</u>	57.3

Table 3: **Speech-to-text translation results (SSA-COMET ↑)**. *Italics* indicate best within each method; **bold** indicates best overall while underlined indicate second best result. Multilingual fine-tuning (FT) is a model fine-tuned across all the Nigerian languages data. Naijá is not supported by NLLB, we label it N/A

			XX → Eng					Eng → XX				
Method		Model	Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
CASCADDED (ASR + MT)	Omnilingual-ASR	+ NLLB	42.0	<u>33.8</u>	N/A	38.1	–	50.0	45.2	N/A	29.1	–
		+ Gemma 4 Zero-Shot	34.7	23.1	35.4	26.9	30.0	36.3	24.9	<u>20.8</u>	11.8	23.5
		+ Gemma 4 Few-Shot	36.9	24.7	36.6	26.1	31.1	<u>37.2</u>	26.8	22.8	11.3	24.5
END-TO-END	SeamlessM4T	Zero-Shot	15.9	19.4	<u>31.1</u>	<u>43.3</u>	27.4	N/A	24.5	N/A	14.2	-
		Mono-FT	41.9	39.2	18.1	43.5	35.7	13.7	54.7	6.1	<u>37.0</u>	27.9
		Multi-FT	38.0	36.0	40.1	41.8	39.0	14.5	55.3	15.3	37.1	<u>30.6</u>

Table 4: **Speech-to-text translation results (ChrF++ ↑)**. **Bold** indicates best overall while underlined indicate second best result. Multilingual fine-tuning (Multi FT) uses all Nigerian data. NLLB does not support Naijá.

speech outputs, thereby eliminating the need for an intermediate ASR. For the S2TT task, we evaluate SeamlessM4T-Large V2 (2.3B) in *three settings*: (1) **Zero-shot inference on the test set** (2) **Monolingual fine-tuning (Mono-FT)** per language to further improve generalization (3) **Multilingual fine-tuning (Multi-FT)** that combines all languages per translation direction (Eng↔low-resource languages (LRL)). For Mono-FT, each language-specific model is trained with a learning rate (LR) of 1e-5, 16 gradient accumulation (GA) steps and 3 epochs. LR of 5e-6, 32 GA steps and 3 epochs is used for Multi-FT.⁴

For LRL↔English S2ST evaluation, we use SeamlessM4T-large 2.3B for both zero-shot and Mono-FT settings, as SeamlessM4T-Large V2 does not support S2ST finetuning. As Hausa is not supported in SeamlessM4T, it is mapped to a proxy language token from the same language family (Arabic), following Adelani et al. (2022a). Similarly, we map Naijá to French, as it is very similar to English and should therefore be linked to a European language. The results for these two languages should therefore be interpreted as a fine-tuned extension rather than a reflection of the model’s native multilingual coverage. Training is conducted with a LR of 1e-5, batch size 2 and a maximum of 10

epochs with early stopping. Note that the other direction in S2ST (i.e. English→LRL) is omitted because SeamlessM4T only supports translation into high-resource languages.

AudioLLMs We evaluate proprietary LLMs that support low-resource languages for S2TT and S2ST tasks (i.e., Gemini 2.5, Gemini 3.1, GPT-Audio). For comparison, we also evaluate Gemma 4 for S2TT as an open-source option. As Gemini does not currently support native S2ST, we construct a cascaded S2ST pipeline by applying Gemini 2.5 TTS to the outputs of its S2TT system. Note that most of the other high-performing AudioLLMs/multimodal LLMs, such as Qwen3-Omni (Xu et al., 2025) or DeSTA2 (Lu et al., 2025) do not support our studied languages.

4.2 Evaluation Metrics

Both lexical and embedding-based metrics are used to obtain a comprehensive assessment of model performance. Specifically, we report SSA-COMET (-MTL) (Li et al., 2025), which is an extension of the COMET (Rei et al., 2020) metric for African languages. COMET incorporates semantic similarity between hypothesis and reference translations by leveraging pretrained multilingual encoders to compute contextualized sentence representations. We additionally use ChrF++ (Popović, 2015), a character n-gram-based metric that computes an

⁴Hyperparameters selected via hyperparameter tuning.

Model	Method	XX → Eng					Eng → XX				
		Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
SeamlessM4T	Mono FT	54.9	<u>52.3</u>	50.7	60.3	54.6	13.7	54.7	6.1	37.0	27.9
	Multi FT	46.7	47.4	56.0	54.3	51.1	53.4	64.6	42.6	68.5	57.3
Gemma 4	Zero-Shot	39.2	31.7	52.7	37.0	40.2	40.6	25.8	53.2	30.9	37.6
	Few-Shot	41.5	28.0	52.9	35.7	39.5	39.3	30.0	51.3	32.2	38.2
GPT-Audio	Zero-Shot	34.0	31.1	60.3	36.0	40.4	38.2	40.2	42.1	40.6	40.3
	Few-Shot	33.8	30.4	60.3	35.4	40.0	63.2	61.3	60.2	61.2	61.5
Gemini 2.5	Zero-Shot	60.9	37.0	59.1	49.4	51.6	63.8	64.6	58.2	67.2	63.5
	Few-Shot	63.1	41.0	62.2	51.2	54.4	64.2	64.4	54.3	68.9	63.0
Gemini 3.1	Zero-Shot	<u>68.1</u>	51.3	<u>65.4</u>	<u>63.6</u>	<u>62.1</u>	<u>68.3</u>	<u>67.1</u>	<u>61.5</u>	<u>72.1</u>	<u>67.3</u>
	Few-Shot	69.4	56.2	66.7	65.9	64.6	68.3	67.4	62.8	72.3	67.7

Table 5: **Speech-to-text translation results (SSA-COMET ↑) for AudioLLM**, compared with fully supervised fine tuning (Seamless M4T). *Italics* indicate best within each method; **bold** is best overall. Second best is underlined.

F-score over overlapping character sequences between the hypothesis and reference. By operating at the character level, ChrF++ is more sensitive to morphological variations and orthographic similarity than word-level metrics, but are sometimes unreliable for many LRLs (Freitag et al., 2022; Wang et al., 2024).

For S2ST metrics, we calculate ASR-COMET by performing ASR on the predicted audio translations using Omnilingual-ASR, followed by SSA-COMET on the transcriptions, as in the S2TT task.

5 Results

5.1 Speech-to-Text Translation Results

Architectural Comparisons Overall, we observe a consistent advantage of E2E and Audio LLM-based approaches over traditional cascaded S2TT pipelines (Tables 3–5). In particular, the best E2E configuration (i.e., SeamlessM4T with Mono-FT or Multi-FT) and the strongest Audio LLM (Gemini 3.1) both surpass the cascaded Omnilingual ASR + MT pipeline. This trend aligns with prior findings that cascaded S2TT systems are susceptible to error propagation, where transcription errors from ASR compound downstream during MT (Etchegoyhen et al., 2022). To assess whether degradation in cascaded performance stems primarily from ASR errors or MT limitations, we pair a fixed Omnilingual ASR backbone with multiple MT models (Tables 3, 4). The results show that for Hausa and Igbo, performance remains relatively similar across NLLB and TinyAya backends regardless of language direction. In contrast, Yorùbá exhibits substantial sensitivity to the choice of MT model: replacing NLLB with Tiny Aya leads to large drops in SSA-COMET (e.g., 50.6 → 35.5

for XX→Eng; 58.0 → 48.4 for Eng→XX).^{5,6}

End-to-End Fine-tuning Table 3 reveals a highly asymmetric zero-shot behavior of SeamlessM4T across directions and languages. In the LRL→English setting, performance fluctuates: while Yorùbá and Naijá results are relatively high at 52.4 and 57.0 SSA-COMET scores, Igbo and Hausa lag significantly behind (20.6 and 14.6, respectively). The particularly low performance for Hausa and Naijá for English→LRL is expected given their unsupported status in the base model.

Inconsistencies of metrics to wrong language output In the reverse English→LRL direction, zero-shot scores for Igbo and Yorùbá appear to exhibit relatively high SSA-COMET scores (53.1 & 55.4, respectively). However, this is not reflected in ChrF++ (Table 4), where scores remain substantially lower. This discrepancy highlights a key limitation of embedding-based metrics: SSA-COMET is more tolerant to lexical and surface-form mismatches, whereas ChrF++ penalizes such deviations more directly. Manual inspection of the outputs shows that a considerable portion of SeamlessM4T predictions remain in the source language. This indicates that the model often fails to fully perform the translation despite achieving moderate semantic similarity scores, and highlights the need to have multiple evaluation metrics. While ChrF++ is generally unreliable as it often achieves low scores for languages like Yorùbá with extensive use of diacritics and penalizes progress (i.e.

⁵Qualitative inspection reveals that Tiny Aya frequently produces extraneous “chain-of-thought”-like tokens and meta-commentary in Yorùbá outputs, suggesting a mismatch between instruction-tuned generative behavior and the constrained requirements of translation.

⁶See Appendix E for results with other metrics.

			XX → Eng				Eng → XX			
Method	Model		Hausa	Igbo	Naijá	Yorùbá	Hausa	Igbo	Naijá	Yorùbá
Naijá-accent										
End-to-End	SeamlessM4T	Multilingual	23.5	26.4	N/A	36.3	N/A	N/A	N/A	N/A
Cascaded	Omni + NLLB/TinyAya	+ Gemini 2.5 TTS Naijá	50.4	40.6	40.8	46.6	37.1	35.0	47.2	39.8
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS Naijá	<u>61.8</u>	<u>50.7</u>	<u>60.0</u>	<u>65.4</u>	45.8	<u>40.0</u>	<u>47.6</u>	42.5
British-accent										
Cascaded	Omni + NLLB/TinyAya	+ Gemini 2.5 TTS British	51.5	41.8	51.4	48.6	40.3	38.3	41.2	<u>45.1</u>
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS British	63.3	51.2	62.4	<u>60.7</u>	<u>44.4</u>	40.4	47.9	51.1

Table 6: **Speech-to-speech translation results (ASR-COMET ↑)**. *Italics* indicate best within each method; **bold** indicates best overall while second best is underlined. Best models per method are used as a base for TTS. Thus, we use Omni+NLLB for Hausa, Igbo and Yorùbá, and Tiny Aya for Naijá, as it is not supported by NLLB.

the language with the highest SSA-COMET score out of the three languages achieved lower score with ChrF++ in Table 4), it is able to catch simpler errors that more sophisticated metrics cannot.

SeamlessM4T exhibits notably different behaviour on Naijá compared to the other language pairs. While Mono-FT improves SSA-COMET scores over the Zero-Shot setting, the opposite trend is observed for ChrF++. One likely factor is the relatively small amount of available Naijá training data, which makes fine-tuning, particularly through the use of a proxy language, more difficult.

Fine-tuning helps wrong language generation

Fine-tuning substantially mitigates these issues, but the gains depend critically on both directionality and training strategy. In the LRL→Eng direction, Multi-FT yields large improvements for Hausa, Igbo and Naijá which suggests that shared language representations help compensate for limited per-language data. Yet, this comes at a slight cost for Yorùbá, where zero-shot performance is already strong, and indicates potential cross-lingual interference. On the other hand, Mono-FT consistently improves performance across all languages and achieves the best overall results in this direction (e.g., 60.3 for Yorùbá), with Naijá being the exception, likely due to the comparatively limited amount of available training data. This suggests that joint training could introduce cross-lingual trade-offs that are absent in language-specific adaptation.

The trend reverses for English→LRL, where Multi-FT consistently outperforms Mono-FT across both SSA-COMET and ChrF for Hausa, Igbo and Yorùbá, and still achieves relatively high results for Naijá. Given the lower-resource nature of this setting, this suggests that performance is more strongly influenced by overall data availability, with shared multilingual training providing

Method	XX → Eng				Eng → XX			
	ha	ig	yo	Avg.	ha	ig	yo	Avg.
CASCADED	<u>75.0</u>	<u>47.5</u>	<u>74.7</u>	<u>65.7</u>	<u>82.1</u>	<u>73.9</u>	<u>76.0</u>	<u>77.3</u>
E2E	4.1	29.0	59.3	30.8	N/A	N/A	N/A	N/A
AUDIOLLM	99.2	72.8	93.2	88.4	88.5	83.4	94.0	88.6

Table 7: **S2ST human evaluation results (0-100 average Likert scale ↑)**. *Italics* indicate best within each method; **bold** is best overall. Underlined is second best.

greater benefit than language-specific fine-tuning.

AudioLLM Evaluation As shown in Table 5, Audio LLMs establish the highest baseline for S2TT, with Gemini 3.1 emerging as the strongest model across all language pairs and directions. Specifically, under few-shot prompting, Gemini 3.1 achieves the best overall SSA-COMET scores, and surpasses both the best fine-tuned SeamlessM4T models (average margin of -8 points) and all cascaded systems. Moreover, it is consistently seen across Gemini models that few-shot prompting helps in providing modest but reliable gains over its zero-shot performance. This indicates that in-context learning helps stabilize cross-lingual alignment in low-resource settings. In contrast, GPT-Audio exhibits weaker and less stable performance particularly for LRL→English, where few-shot prompting does not consistently improve results, and even leads to degradation. Similarly, Gemma 4 does not improve with few-shot prompting, and reports the lowest average scores with the AudioLLM category, and is lower than that of SeamlessM4T. See Appendix A.2 for more information.

5.2 Speech-to-Speech Translation Results

Speech-to-speech systems Table 6 presents S2ST results across cascaded, E2E, and AudioLLM-based approaches. The findings reinforce a consistent pattern observed in S2TT; performance is largely driven by the quality of the

translation component rather than speech generation. For LRL→English, AudioLLM pipelines achieve the strongest results overall, with Gemini 3.1 (few-shot) + TTS reaching the highest score for Hausa (63.3), outperforming cascaded systems by a 12.9 margin. This advantage is reflected across the other languages, all outperforming the cascaded methods by a margin over 10. Note that accent variation in the TTS component (Received Pronunciation (RP) British vs. Nigerian English) has minimal impact in the LRL→English direction, with only marginal differences across all languages. The primary exception is the cascaded approach on Naijá, where performance decreases noticeably under British-accented speech. This may be explained by the linguistic similarity between Naijá and Nigerian-accented English, which likely provides an advantage when the target speech aligns more closely with the source language context. Overall, these findings suggest that accent choice in the TTS component has limited influence on speech-to-speech quality when English is the target language.

In contrast, the English→LRL direction reveals a clearer advantage for AudioLLM-based approaches compared to cascading systems. Gemini 3.1 with Nigerian English (Naija) TTS achieves the best performance across most of the target languages (e.g., 45.8 for Hausa, 40.0 for Igbo and 47.6 for Naijá), while the cascaded approach remains more effective for Yorùbá (39.8). Furthermore, S2ST with SeamlessM4T (multilingual) performs significantly worse than both cascaded and AudioLLM approaches, which is likely due to weaker translation quality from the V1 model, and simply outputted English for Naijá. This highlights that translation quality remains the primary bottleneck in S2ST performance.

5.3 Human Evaluation

To assess translation quality, we conduct six human evaluation studies, corresponding to each English↔XX language pair. Each survey is completed by three bilingual annotators, who evaluate both the semantic adequacy of the translation and the pronunciation quality of the generated S2ST output on a 0–100 scale. For 100 randomly sampled utterances from the Nigerian-context test set, we construct one evaluation item per modeling paradigm, using the strongest system in each: Omnilingual-ASR + NLLB (cascaded), SeamlessM4T Multi-FT (E2E), and Gemini 3.1 +

Gemini 2.5-TTS (AudioLLM). Since Eng→XX lacks an E2E paradigm, we evaluate only the cascaded and AudioLLM settings, yielding 200 items for Eng→XX and 300 for XX→Eng. Further details can be found in Appendix H.

Overall, the human evaluation results reveal a consistent hierarchy across language pairs and translation directions. AudioLLM systems obtain the highest translation quality ratings, with average scores exceeding 88%, whereas cascaded systems achieve moderate performance (65–77%) and E2E models receive the lowest ratings (Table 7). Across the evaluated languages, Igbo receives the lowest human scores, mirroring the pattern observed in the automatic evaluation results in Table 6. This outcome may reflect Igbo’s comparatively lower-resource status relative to the other languages.

6 Further Analysis

6.1 Sentence Length Distribution

To further investigate translation quality, we analyze prediction-to-ground-truth sentence length ratios across architectures. Figures 2 and 3 show these distributions for the LRL→Eng and Eng→LRL directions, respectively. The cascaded approach yields the most stable and compact distribution, centering tightly around 1.0 with minimal variance. Although cascaded systems can propagate downstream ASR errors, their intermediate bottleneck helps enforce structural alignment and prevents runaway generation or severe truncation.

Conversely, SeamlessM4T shows severe instability, with bottom-heavy distributions that indicate systematic under-generation, especially for Hausa. This likely stems from Hausa being unsupported in the base SeamlessM4T model and handled through proxy Arabic tokens, leading to frequent generation failures, untranslated source-language outputs and repetitive words.

Audio LLMs exhibit the highest variance, with upper tails frequently exceeding a ratio of 2.0, indicating over-generation. Manual inspection of the 50 longest over-generated outputs revealed that all were valid, verbose paraphrases that unpack compact clauses into explicit compound sentences while preserving meaning. This accounts for the discrepancy between human preference for Audio LLM outputs and their apparent over-generation, as sentence-length ratio metrics penalize such verbosity, whereas human evaluators and semantic metrics (e.g., SSA-COMET) rate it positively, as translation accuracy remains unaffected. Further

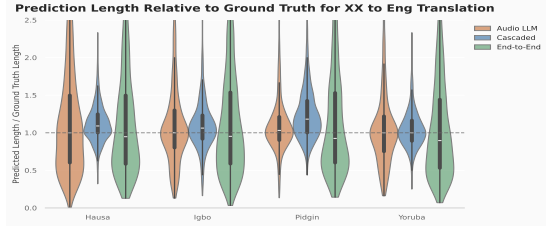


Figure 2: Prediction and ground truth sentence length ratio for $XX \rightarrow$ English translation.

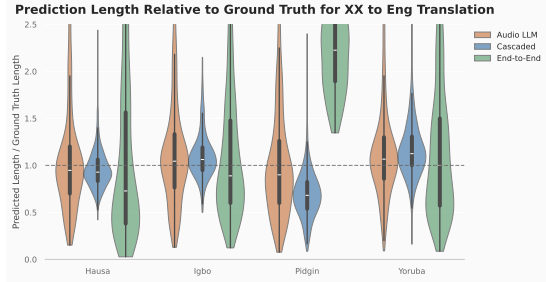


Figure 3: Sentence length ratio for English $\rightarrow$ XX .

analyses are included in Appendix G.

6.2 Part-of-Speech (POS) Analysis

To further understand the failure modes of different paradigms, we conduct POS error analysis on the outputs of best-performing AudioLLM and cascaded systems (Figure 4 and Appendix F). POS tagging is performed automatically using open-source taggers, which are monolingually fine-tuned AfroXLMR-76L (Adelani et al., 2024) encoder models that are trained on MasakhaPOS (Dione et al., 2023). Analysis is restricted to $Eng \rightarrow XX$, tagging both reference and predicted target-language text (with the same tagger) to isolate target-generation quality.

It can be observed that aggregate scores mask systematic paradigm differences. POS-level breakdown shows AudioLLM errors concentrate on open-class, content-bearing tags (e.g., X, ADV, VERB, SYM) and are lowest on closed-class tags (e.g., CCONJ, PROPN, NUM). Cascaded systems show the opposite pattern, with dominant SYM, NUM, and PUNCT, and VERB remaining comparatively stable. We attribute this to differing inductive biases; AudioLLMs are prone to semantic drift on content words, while cascaded systems are more vulnerable on transcription-sensitive categories due to the intermediate textual bottleneck.

Moreover, Yorùbá diverges from the other languages, showing elevated SYM and INTJ errors and a flatter error distribution, whereas Hausa, Igbo,

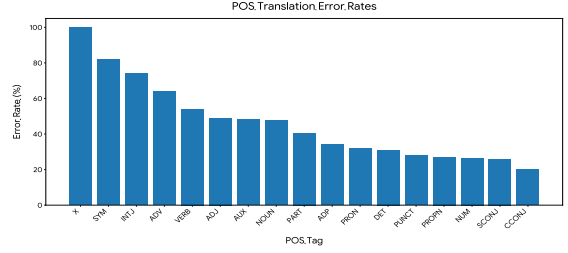


Figure 4: POS error rates for English $\rightarrow$ Yorùbá S2TT using best-performing AudioLLM.

and Naijá show broadly similar rankings across paradigms. This aligns with Yorùbá’s greater sensitivity to MT-model choice (Sec 5.1), likely reflecting its profile as a highly isolating language reliant on tone marks and underdots (Ògúnremí et al., 2024). Since Yorùbá encodes POS prosodically and orthographically rather than morphologically, these distinctions are fragile under S2T transcription and harder for aggregate metrics to capture (see Appendix F).

7 Conclusion

In this paper, we introduced **NaijaS2ST**, a parallel speech–text dataset for Igbo, Hausa, Yorùbá, and Nigerian Pidgin paired with English, designed to address the scarcity of high-quality, diverse speech resources for low-resource languages in African contexts. The dataset enables standardized evaluation of both S2TT and S2ST systems under realistic multi-accent and multilingual conditions, with carefully curated recordings spanning diverse speakers, dialects, and quality-controlled splits.

Using NaijaS2ST, comprehensive benchmarking studies were conducted across cascaded, E2E, and AudioLLM-based approaches in bidirectional translation settings. Results demonstrate a clear advantage of E2E and AudioLLM systems over traditional cascaded methods, with AudioLLMs achieving the strongest overall performance. Fine-tuning is essential for E2E models, but exhibits strong directionality effects; monolingual adaptation is most effective for $LRL \rightarrow$ English, whereas multilingual training yields greater gains in the reverse direction. Moreover, translation quality remains the primary bottleneck for S2ST, while speech synthesis choices including accent variation, have comparatively limited impact in $LRL \rightarrow$ English scenarios. We hope this work will support future research toward more robust, scalable, and inclusive multilingual speech translation research.

8 Limitations

Despite the breadth of our study, several limitations remain. To start, our evaluations are conducted in a controlled offline setting and therefore does not capture practical deployment considerations such as inference latency, streaming constraints, or computational efficiency. In particular, although AudioLLM-based systems achieve strong translation performance, they may incur substantially higher inference cost and latency than cascaded or E2E approaches, potentially limiting their applicability in real-time or resource-constrained environments. Moreover, while we have benchmarked across varying cascaded, E2E, and AudioLLM-based paradigms, our exploration of model configurations is not exhaustive. For AudioLLMs especially, we evaluate only a limited set of prompting and decoding strategies, and do not systematically explore the broader design space of prompt engineering, in-context learning configurations, or alignment techniques. Given the sensitivity of AudioLLMs to prompt formulation, further performance gains may be achievable through more extensive optimization.

9 Ethical considerations

This project involved collecting speech data from more than 250 participants across Nigeria. All participants provided their informed consent for the use of their recordings in speech translation research and were fairly compensated for their contributions. Given the scale of the data collection effort, we did not rely on volunteer recordings. Instead, each participant recorded approximately 250 sentences and received \$15 in compensation (which is a fair rate in Nigeria). Similarly, language coordinators and translators were also adequately compensated for their work. Translators were paid \$0.50 per sentence, resulting in a total translation cost exceeding \$11,000 in all languages. Throughout the project, we ensured that contributors were compensated fairly while also allowing room for negotiation when necessary.

Regarding privacy and ethics, we did not collect private, personal, or sensitive content, nor did we ask participants to read such material. All recorded sentences were sourced from publicly available corpora developed for prior NLP research projects, including NTREX, SSA-MT, and VOA datasets. To further protect participant privacy, we will prohibit the use of the dataset for tasks intended to

re-identify speakers. In addition, we do not intend to release any personally identifiable information (PII) associated with participants, such as names, email addresses, phone numbers, or other identifying metadata.

10 Acknowledgments

This research was supported by the Google grant via Mila, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and in part by the AI2050 program at Schmidt Sciences. This work was partially supported by Azure sponsorship credits granted by Microsoft’s AI for Good Research Lab and Google’s Gemini Academic Program Award. We thank Elizabeth Salesky for her careful guidance and thoughtful feedback throughout this research. Finally, we thank Luel Hagos Beyene for his technical assistance with the speech-to-text model fine-tuning.

References

- Ife Adebbara, Oluwaseun Nifemi, Rashidat Damilola Sikiru, Olanrewaju Israel Lawal, Ololade Anjuwon, Olubayo Adekanmbi, Anthony Soronnadi, John Emeka Eze, and Ewezu Ngim Ngim. 2026. [African voices nigeria: 2500 hours of ethically sourced speech data for four nigerian languages](#). In *7th Workshop on African Natural Language Processing*.
- David Ifeoluwa Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, and 1 others. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070.
- David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022b. [A few thousand translations go a long way! leveraging pre-trained models for African news translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.
- David Ifeoluwa Adelani, A. Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. 2025. [Does generative AI speak Nigerian-Pidgin?: Issues about repre-](#)

- sentativeness and bias for multilingualism in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1571–1583, Albuquerque, New Mexico. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-’Eric Saint-James, Yiannis (Ioannis) Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere-Lluís Huguet Cabot, João Janeiro, Jean Maillard, Gabriel Mejia Gonzalez, and 11 others. 2026. [Omnilingual mt: Machine translation for 1,600 languages](#). *ArXiv*, abs/2603.16309.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: a framework for self-supervised learning of speech representations](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, and 49 others. 2025. [Joint speech and text machine translation for up to 100 languages](#). *Nature*, 637(8046):587–593.
- Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. [Cascade versus direct speech translation: Do the differences still make a difference?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.
- Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. [Speech-to-speech translation for a real-world unwritten language](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.
- Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan van Esch, Vera Axelrod, Simran Khanuja, Jonathan Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, and Melvin Johnson. 2022a. [XTREME-S: Evaluating Cross-lingual Speech Representations](#). In *InterSpeech 2022*, pages 3248–3252.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022b. [Fleurs: Few-shot learning evaluation of universal representations of speech](#). *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.
- Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Bar-rault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630:841 – 846.
- Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, and 25 others. 2023. [MasakhaPOS: Part-of-speech tagging for typologically diverse African languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.
- Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. [Leveraging Pseudo-labeled Data to Improve Direct Speech-to-Speech Translation](#). In *Interspeech 2022*, pages 1781–1785.
- Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2023. [SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers*), pages 16251–16269, Toronto, Canada. Association for Computational Linguistics.
- Chris Emezue, The NaijaVoices Community, Busayo Awobade, Abraham Toluase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. [The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages](#). In *Interspeech 2025*, pages 1338–1342.
- Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G. Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. 2022. [Cascade or direct speech translation? a case study](#). *Applied Sciences*, 12(3).
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. [CVSS corpus and massively multilingual speech-to-speech translation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.
- Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. [Direct speech-to-speech translation with a sequence-to-sequence model](#). In *Proc. Interspeech 2019*, pages 1123–1127.
- Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022a. [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, Dublin, Ireland. Association for Computational Linguistics.
- Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. [Textless speech-to-speech translation on real data](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.
- Senyu Li, Jiayi Wang, Felermimo D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenertorp, and David Ifeoluwa Adelani. 2025. [SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12979–12998, Suzhou, China. Association for Computational Linguistics.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-Yi Lee. 2025. [Developing instruction-following speech language model without speech instruction-tuning data](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon KABONGO KABENAMUALU, Elizabeth Salesky, Iro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete AGBOLO, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. [BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus](#). In *Interspeech 2022*, pages 2383–2387.
- Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayuth Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2024. [Translatotron 3: Speech to speech translation with monolingual data](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10686–10690.
- Palanichamy Naveen and Pavel Trojovský. 2024. [Overview and challenges of machine translation for contextually appropriate translations](#). *iScience*, 27(10):110878.
- Tolúloṣe Ògúnṛémí, Kólá Túbòsún, Anuoluwapo Aremu, Iro Orife, and David Ifeoluwa Adelani.

2024. [ÌròyìnSpeech: A multi-purpose Yorùbá speech corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9296–9303, Torino, Italia. ELRA and ICCL.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. [AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR](#). *Transactions of the Association for Computational Linguistics*, 11:1669–1685.
- 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, Lisbon, Portugal. Association for Computational Linguistics.
- Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. [Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation](#). In *Interspeech 2022*, pages 5195–5199.
- Qian qian Dong, Zhiying Huang, Qiao Tian, Chen Xu, Tom Ko, yunlong zhao, Siyuan Feng, Tang Li, Kexin Wang, Xuxin Cheng, Fengpeng Yue, Ye Bai, Xi Chen, Lu Lu, Zejun MA, Yuping Wang, Mingxuan Wang, and Yuxuan Wang. 2024. [Polyvoice: Language models for speech to speech translation](#). In *The Twelfth International Conference on Learning Representations*.
- 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.
- Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Ryan Padfield, James Qin, Daniel Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle D. Tadmor, Ramanovich, and 12 others. 2023. [Audiopalm: A large language model that can speak and listen](#). *ArXiv*, abs/2306.12925.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaei, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. [Tiny aya: Bridging scale and multilingual depth](#). *Preprint*, arXiv:2603.11510.
- Mohammad Sarim, Saim Shakeel, Laeeba Javed, Mohammad Nadeem, and 1 others. 2025. Direct speech to speech translation: A review. *arXiv preprint arXiv:2503.04799*.
- Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ifeoluwanimi Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. [Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages](#). *ArXiv*, abs/2511.09690.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. [VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.
- Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. Covost 2 and massively multilingual speech translation. In *Interspeech*, volume 2021, pages 2247–2251.
- Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehghoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. [AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.
- Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021. Uwspeech: Speech to speech translation for unwritten languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14319–14327.
- Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. [Stream-Speech: Simultaneous speech-to-speech translation](#)

with multi-task learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8964–8986, Bangkok, Thailand. Association for Computational Linguistics.

A British English Dataset Addition

A.1 Dataset

Using the same English data from our accented-English dataset, we record a new British English test set and synthesize the train set using a mixture of 10 Gemini 2.5 TTS voices. We provide our own recording tool to the British annotators, who were instructed the speakers to record on their laptops with a good microphone in a quiet environment. We recruited one female and three males speakers through our community, who are native Southern British English speakers. Each volunteer recorded 250 utterances, covering the 1000 necessary for the dev and test sets. The speakers were aware that their recordings would be made public, but used only for research purposes, as were the Nigerian languages participants.

A.2 Experiments with Accents

We further investigate the difference in performance between Nigerian and British accented English, the best cascaded and AudioLLM methods are evaluated (Table 5) on the newly created dataset mentioned in Section A.1.

The results in Table 8 highlight clear differences in robustness to British- and Naijá-accented English inputs across model architectures. The strongest cascaded approaches are substantially affected by accent variation, with performance generally improving under British-accented input compared to Naijá-accented speech. In contrast, the best-performing AudioLLM remains comparatively stable across accent conditions, achieving similar performance for Hausa, Igbo, and Naijá translation under both input varieties. The primary exception is Yorùbá, where performance decreases notably for British-accented input despite achieving the strongest overall results under Naijá-accented speech. This finding suggests that accent alignment between the source speech and target language context may particularly benefit Yorùbá translation. Additionally, the performance gap between zero-shot and few-shot prompting is smaller for British-accented inputs, indicating that in-context examples provide comparatively limited gains in this setting. Overall, these results suggest

that end-to-end AudioLLM approaches exhibit substantially greater robustness to accent variation than cascaded ASR+MT pipelines.

Model	Method		Eng → XX				
			Hausa	Igbo	Naijá	Yorùbá	Avg.
British-accent CASCADED (ASR + MT)	Omnilingual-ASR	+ NLLB	52.6	58.4	10.9	63.2	46.3
		+ Tiny Aya	49.3	53.2	45.4	44.9	48.2
Naijá-accent CASCADED (ASR + MT)	Omnilingual-ASR	+ NLLB	47.6	52.6	9.7	58.0	42.0
		+ Tiny Aya	39.0	40.0	15.7	26.3	30.3
British-accent AUDIOLLM	Gemini 3.1	Zero-Shot	68.2	67.2	<u>62.7</u>	63.1	65.3
		Few-Shot	68.9	67.7	61.8	63.0	65.4
Naijá-accent AUDIOLLM	Gemini 3.1	Zero-Shot	<u>68.3</u>	67.1	61.5	<u>72.1</u>	<u>67.3</u>
		Few-Shot	<u>68.3</u>	<u>67.4</u>	62.8	72.3	67.7

Table 8: **Speech-to-text translation results (SSA-COMET ↑) for AudioLLM and Cascade evaluation** for British English → XX. *Italics* indicate best within each method; **bold** indicates best overall while the second best is underlined.

Speech-to-Speech Evaluation Omnilingual-ASR 1B model is used to transcribe speech outputs, on which SSA-COMET and ChrF++ scores are computed. We also finetune Omnilingual on our train and dev sets, to create Naijá-Omni. To analyze the impact of the ASR system on the evaluation, we also synthesize the text translation to RP British English, and compare results between British and Naijá-accented speech and the base Omnilingual and Naijá-Omni. As can be seen in Table 9, Naijá-Omni generally improves results both for the Nigerian and the British accented speech translations. Interestingly, British TTS still gets better results than Naijá TTS using Naijá-Omni. These results indicate that biases introduced at the TTS stage propagate into evaluation metrics, motivating further investigation into accent-invariant evaluation strategies for S2ST.

B Recording Instructions

Participants were instructed to record in quiet environments, speak clearly, and avoid word repetitions. Following the initial recordings, feedback was provided on audio quality, along with guidance on segments requiring re-recording.

C License

All resources released as part of this work are distributed under the CC-BY license.

Model	Output Accent	ASR model	XX → Eng			
			Hausa	Igbo	Naijá	Yorùbá
Gemini 3.1	Naijá	+ Naijá-Omni	61.8	50.7	60.0	65.4
		+ Omni	60.8	50.0	60.7	58.6
	British	+ Naijá-Omni	63.2	51.4	62.6	60.3
		+ Omni	63.3	51.2	62.4	60.7

Table 9: S2ST ASR-COMET ($\uparrow$) results computed with finetuned Naijá-Omni ASR and the original Omnilingual-LLM-1B (Omni).

D Prompt Templates

```
[PROMPT CONTENTS]
You are a translation expert. Here are
{len(fewshot_files)} examples of
{language} audio transcribed, then
translated into English. Following
these examples, transcribe the last
given audio, and use the
transcription to provide its exact
English translation. Return only the
English translation without any
additional text.
[AUDIO] .wav
```

Figure 5: Prompt template used for LRL→English speech-to-text translation in Gemini 2.5 and Gemini 3.1.

```
[USER]
You are a professional translator. Here
are {number_of_few_shot_examples}
examples of {language}
transcriptions and their English
translations:\n{examples_str}\n\n"
Following these examples, translate the
following {language} transcription
to English.
Only output the English translation
without any additional text or
formatting.{input_text}
```

Figure 6: Prompt template used for LRL → Eng MT for Tiny Aya.

```
[SYSTEM]
You are a translation expert. Here are
{len(fewshot_examples)} examples of
{language} audio transcribed, then
translated into English. Following
these examples, transcribe the last
given audio, and use the
transcription to provide its exact
English translation. Return only the
English translation without any
additional text."
[USER]
Translate this {language} audio into
English.
[AUDIO]: .wav
```

Figure 7: Prompt template used for LRL→English speech-to-text translation for GPT-Audio

E ChrF++ and SpBLEU Results

In addition to the SSA-COMET results reported previously, we also provide ChrF++ scores for other S2TT experiments. Furthermore, we include **SpBLEU**, which is a metric designed for machine translation that evaluates translation quality. We also report a variant, **ASR-SpBLEU**, which applies the same evaluation procedure on ASR-transcribed hypotheses, thereby capturing the compounded effects of recognition and translation errors in cascaded or speech-based pipelines. Table 10 through Table 14 explicitly detail the comprehensive SpBLEU, ChrF++, and ASR-SpBLEU metrics. They serve to provide a highly granular breakdown of the evaluation settings discussed in Section 5, showing the performance variations across both Naijá-accented and British-accented evaluations.

Method	Model		XX → Eng					Eng → XX				
			Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
CASCADED (ASR + MT)	Omnilingual-ASR	+ NLLB	<u>17.3</u>	11.0	16.1	17.0	11.1	16.7	29.2	0.0	23.4	17.32
		+ Tiny Aya	12.4	6.2	4.3	10.9	8.45	<u>16.2</u>	23.8	1.1	18.7	15.0
END-TO-END	SeamlessM4T	Zero-Shot	1.3	4.0	N/A	19.5	12.4	N/A	7.5	N/A	2.5	5.0
		Mono FT	18.6	17.6	1.2	21.1	14.7	1.1	<u>37.2</u>	0.3	<u>30.0</u>	22.9
		Multi FT	13.9	<u>14.8</u>	18.5	15.8	15.8	1.5	40.0	1.2	30.3	23.8

Table 10: **Speech-to-text translation results (SpBLEU ↑)**. *Italics* indicate best within each method; **bold** indicates best overall while underlined indicate second best result. Multilingual fine-tuning (FT) is a model fine-tuned across all the Nigerian languages data.

Model	Method	XX → Eng					Eng → XX				
		Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
SeamlessM4T	Zero-Shot	1.3	4.0	7.4	8.3	5.3	N/A	7.5	N/A	2.5	-
	Mono FT	18.6	17.6	1.2	21.1	14.7	1.1	<u>37.2</u>	0.3	30.0	17.2
	Multi FT	13.9	14.8	18.5	15.8	16.1	1.5	40.0	1.5	30.3	18.3
Gemini 2.5	Zero-Shot	19.2	7.2	12.5	13.5	13.2	26.7	37.4	5.7	31.2	25.3
	Few-Shot	23.3	10.9	16.5	17.0	16.9	36.1	5.2	25.2	29.1	23.9
Gemini 3.1	Zero-Shot	<u>30.0</u>	<u>19.6</u>	<u>28.4</u>	21.1	<u>24.8</u>	<u>30.3</u>	39.0	6.6	<u>35.6</u>	<u>27.9</u>
	Few-Shot	35.6	25.2	33.3	28.4	38.7	32.4	40.6	<u>20.4</u>	36.3	32.4
GPT-Audio	Zero-Shot	4.4	4.6	16.2	<u>24.7</u>	12.5	7.2	8.1	1.9	6.5	5.9
	Few-Shot	4.3	4.4	16.3	5.6	7.7	25.2	29.8	5.8	13.6	18.6

Table 11: **Speech-to-text translation results (SpBLEU ↑) for AudioLLM evaluation**, with a comparison to fully supervised fine tuning (SFT) (Seamless M4T). **Bold** indicates best overall while the second best is underlined.

Model	Method	XX → Eng					Eng → XX				
		Hausa	Igbo	Naijá	Yorùbá	Avg.	Hausa	Igbo	Naijá	Yorùbá	Avg.
SeamlessM4T	Zero-Shot	15.9	19.4	31.1	43.3	27.4	N/A	24.5	N/A	14.2	-
	Mono-FT	41.9	39.2	18.1	43.5	35.7	13.7	54.7	6.1	37.0	27.9
	Multi-FT	38.0	36.0	40.1	41.8	39.0	14.5	55.3	15.3	37.1	30.6
Gemini 2.5	Zero-Shot	46.5	30.0	38.9	37.4	38.2	49.3	53.1	28.5	32.4	40.8
	Few-Shot	49.3	33.5	41.2	39.4	40.9	48.9	53.8	27.0	37.6	41.8
Gemini 3.1	Zero-Shot	<u>54.9</u>	<u>41.5</u>	<u>44.9</u>	<u>50.8</u>	<u>48.0</u>	<u>53.0</u>	<u>56.3</u>	<u>28.8</u>	<u>40.7</u>	<u>44.7</u>
	Few-Shot	58.1	46.0	49.7	54.5	52.1	53.6	57.0	41.0	42.0	48.4
Gemma 4	Zero-Shot	17.0	15.6	29.9	17.9	20.1	33.2	21.8	22.0	12.8	22.5
	Few-Shot	24.7	15.9	30.2	18.0	22.2	33.5	22.1	22.2	13.0	22.7
GPT-Audio – TO-DO	Zero-Shot	24.2	22.3	39.9	24.7	27.8	23.9	22.7	18.9	14.1	19.9
	Few-Shot	24.1	22.0	40.1	24.4	27.7	48.1	47.2	28.0	24.1	36.9

Table 12: **Speech-to-text translation results (ChrF++ ↑) for AudioLLM evaluation**, with a comparison to fully supervised fine tuning (SFT) (Seamless M4T). **Bold** indicates best overall while the second best is underlined.

			XX → Eng				Eng → XX			
Method	Model		Hausa	Igbo	Naijá	Yorùbá	Hausa	Igbo	Naijá	Yoruba
Naijá-accent										
End-to-End	SeamlessM4T	Multilingual	1.2	1.9	N/A	4.4				
Cascaded	Omni + NLLB/TinyAya	+ Gemini 2.5 TTS Naijá	11.9	<u>7.8</u>	0.1	<u>9.5</u>	11.0	<u>6.4</u>	0.3	<u>12.8</u>
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS Naijá	<u>18.8</u>	<u>12.4</u>	0.3	32.7	13.6	6.6	<u>1.2</u>	14.2
British-accent										
Cascaded	Omni + NLLB/TinyAya	+ Gemini 2.5 TTS British	12.6	8.5	0.4	10.5	11.4	4.6	0.2	9.6
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS British	19.3	12.7	14.1	<u>17.7</u>	12.2	5.6	2.1	<u>14.0</u>

Table 13: **Speech-to-speech translation results (ASR-SpBLEU ↑)**. **Bold** indicates best overall while the second best is underlined.

				XX → Eng				Eng → XX			
Method	Model			Hausa	Igbo	Naijá	Yorùbá	Hausa	Igbo	Naijá	Yoruba
Naijá-accent											
End-to-End	SeamlessM4T	Multilingual		21.4	22.5	N/A	27.0				
Cascaded	Omni + NLLB/TinyAya	+ Gemini 2.5 TTS Naijá		41.7	34.5	14.08	36.3	35.7	26.0	23.9	20.8
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS Naija		47.9	38.0	18.7	53.7	40.5	27.3	22.1	21.6
British-accent											
Cascaded	Omni + NLLB	+ Gemini 2.5 TTS British		38.2	31.8	19.7	34.1	35.4	24.2	13.3	20.1
AudioLLM	Gemini 3.1 Few-Shot	+ Gemini 2.5 TTS British		47.6	37.8	40.3	44.7	37.0	25.5	22.3	26.3

Table 14: **Speech-to-speech translation results (ASR-ChrF ↑)**. **Bold** indicates best overall while the second best is underlined.

F POS Tags

F.1 POS Tagging Error Rate

Since references and predictions may differ in word choice, order, or count, we align their tag sequences using Sequence Matcher-based token alignment rather than assuming 1:1 correspondence, and compute a per-tag error rate:

$$\text{Error Rate}_p = \frac{S_p + D_p}{M_p + S_p + D_p} \times 100 \quad (1)$$

, where M_p , S_p and D_p are match, substitution, and deletion counts for tag p . This method highlights tag-level patterns (e.g. high PROP error rates indicating dropped or mistranslated names) attributable to translation behavior rather than tagging noise.

F.2 Yorùbá POS Tagging Analysis

Because Yorùbá lacks the extensive morphological inflection found in agglutinative languages, syntactic roles (POS) are heavily dependent on exact lexical identification, which in turn is strictly dictated by tones. For example, a minor ASR transcription error in tone can change the word *tí* (a preposition/ADP meaning 'from' or 'of') into *tì* (a verb meaning 'to push' or 'to lock'), or *bí* (a conjunction/SCONJ meaning 'if') into *bì* (a verb meaning 'to ask').

Therefore, when speech models fail to capture these phonological nuances (a common vulnerability in S2T systems), it does not merely result in orthographic typos; it leads to severe lexical substitutions that fundamentally alter the grammatical structure (POS) of the generated sentence.

G Sentence Analysis Ablations

To address the over-generation in AudioLLMs, we ran a quantitative breakdown across all eight translation directions, measuring word-count differences between predictions and references. While over-generation is consistent, its impact is generally low in that the average excess length ranges from 2.46 to 4.19 words per sentence (Table ??), with minimum differences as small as one word. Larger deviations (15-52 words), driven by meta-commentary, "reasoning-like" continuations or more verbose outputs, such as the Igbo→English example below, are outliers, occurring in only a small fraction of cases relative to near-reference-length outputs.

An Example from the Test Set

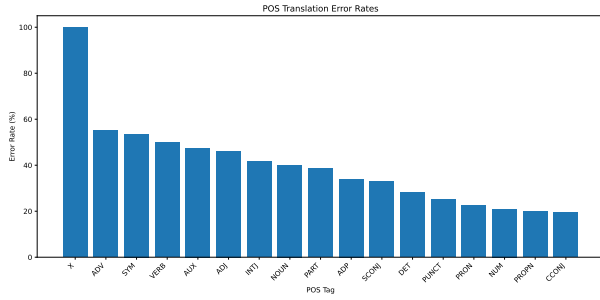
Direction	Samples	Min.	Max.	Avg.
Eng → Hausa	167/500	1	15	2.53
Eng → Igbo	288/500	1	15	3.39
Eng → Naijá	124/500	1	30	3.41
Eng → Yorùbá	333/500	1	16	4.19
Hausa → Eng	765/1500	1	20	3.19
Igbo → Eng	711/1500	1	52	3.95
Naijá → Eng	767/1500	1	20	3.49
Yorùbá → Eng	554/1500	1	9	2.46

Table 15: AudioLLM (Gemini 3.1) over-generation statistics. Difference is defined as the prediction word count minus the reference word count. **Samples** refers to the total number of "over-generated" samples, Min. to the minimum difference, Max. to the maximum difference and Avg. to the average word count difference.

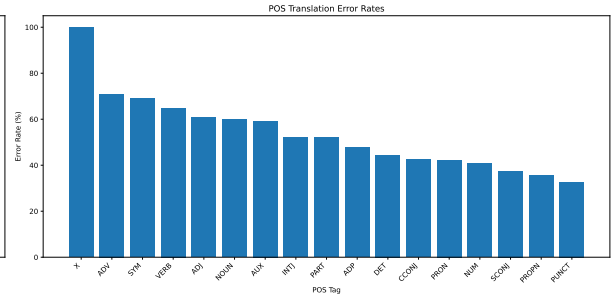
Prediction (31 words): It is the home of golf. The Old Course of the Royal and Ancient Golf Club is located by the ocean. Some of the most famous images in the sporting world.

Reference (ground truth) (7 words): St. Andrews is famous for many things.

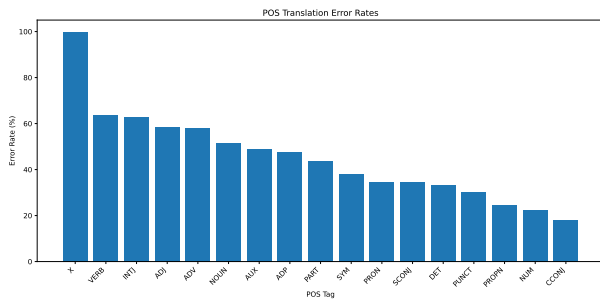
This suggests the length variance is driven by a long tail of infrequent, severe over-generations rather than systematic verbosity, and that the strong automatic/human scores are broadly consistent with generation behaviour.



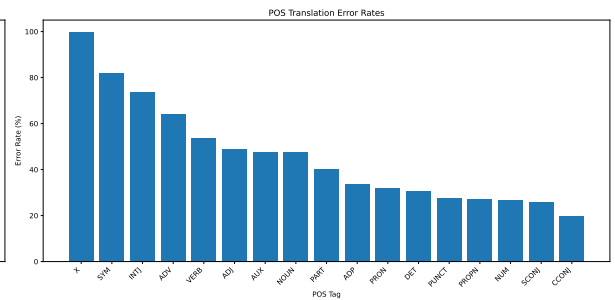
(a) Hausa→English (AudioLLM)



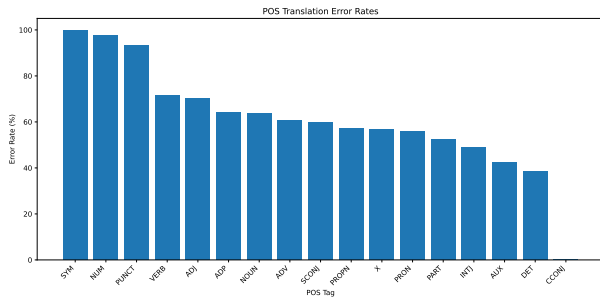
(b) Igbo→English (AudioLLM)



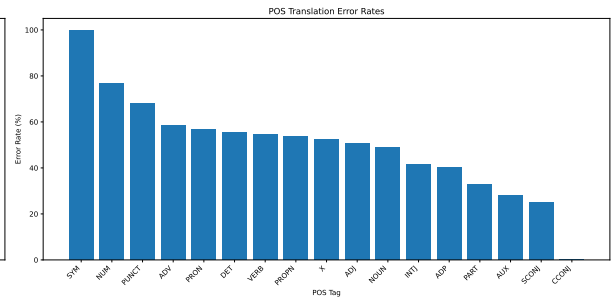
(c) Naijá→English (AudioLLM)



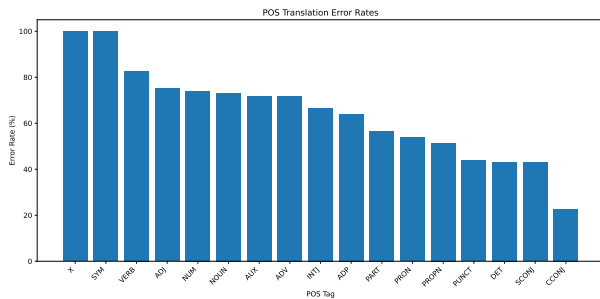
(d) Yorùbá →English (AudioLLM)



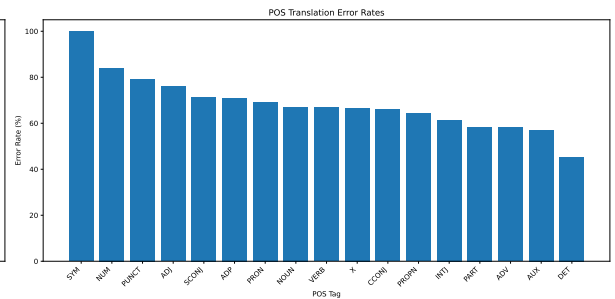
(e) English→Hausa (Cascaded)



(f) English→Igbo (Cascaded)



(g) English→Naijá (Cascaded)



(h) English→Yorùbá (Cascaded)

Figure 8: POS translation error rates across language pairs for AudioLLM (top four) and Cascaded (bottom four) models.

H Human Evaluation Ablations

H.1 Pronunciation Quality

As with translation quality in Section 5.3, a similar trend is observed for pronunciation quality (Table 16). AudioLLMs again achieve the best scores overall, followed by Cascaded systems, while End-to-End models consistently perform worst. These results suggest that AudioLLM approaches produce both more accurate translations and more natural speech outputs across low-resource language settings.

Method	XX → Eng				Eng → XX			
	Hausa	Igbo	Yorùbá	Avg.	Hausa	Igbo	Yorùbá	Avg.
CASCADED (ASR + MT)	97.3	59.3	88.4	81.7	77.1	62.4	55.9	65.1
END-TO-END	78.6	41.4	72.3	64.1	N/A	N/A	N/A	
AUDIOLLM	98.5	69.8	93.6	87.3	83.4	78.6	58.5	73.5

Table 16: S2ST **pronunciation** human evaluation results (0-100 average Likert scale ↑).

H.2 Source Text Addition

To examine the impact of source-text availability, we further duplicate the 300 evaluation items and augment the duplicated versions with the corresponding source transcription, resulting in 600 items in total. In other words, annotators first evaluated 300 target–source audio pairs and later evaluated the same 300 pairs with additional source text provided. We find that for all languages and directions, adding source text had little effect on annotator scores, as shown in Tables 17 ~ 22.

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	85.29	9.54	87.67
	With text	105	85.19	7.40	88.00
Cascaded	No text	105	71.12	13.96	70.67
	With text	105	74.78	11.60	73.33

Table 17: Audio Quality by method and text condition (English to Hausa).

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	68.94	9.65	69.67
	With text	105	73.70	8.36	75.67
Cascaded	No text	105	67.66	8.66	68.33
	With text	105	71.66	8.37	72.00

Table 18: Audio Quality by method and text condition (English to Igbo).

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	76.31	10.47	77.33
	With text	105	77.31	8.52	77.33
Cascaded	No text	105	58.89	16.52	61.33
	With text	105	54.56	15.23	52.67

Table 19: Audio Quality by method and text condition (English to Yorùbá).

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	99.54	2.85	100.00
	With text	105	98.79	5.48	100.00
Cascaded	No text	105	78.70	27.05	91.67
	With text	105	71.24	35.19	93.33
End-to-End	No text	105	6.50	16.08	0.00
	With text	105	1.75	7.81	0.00

Table 20: Audio Quality by method and text condition (Hausa to English).

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	67.39	14.48	70.33
	With text	105	76.55	13.24	81.67
Cascaded	No text	105	63.47	16.07	66.33
	With text	105	64.37	18.11	69.67
End-to-End	No text	105	48.18	17.37	46.33
	With text	105	48.87	17.01	45.67

Table 21: Audio Quality by method and text condition (Igbo to English).

Method	Has Source Text	N	Mean	SD	Median
AudioLLM	No text	105	87.14	11.09	90.67
	With text	105	90.44	11.07	94.67
Cascaded	No text	105	70.96	17.76	73.33
	With text	105	66.14	24.37	69.33
End-to-End	No text	105	44.53	12.58	40.00
	With text	105	36.72	12.42	31.67

Table 22: Audio Quality by method and text condition (Yorùbà to English).

I Annotator Reliability and Agreement

Recall in Appendix H.2 that there were 300 target-source audio pairs, and an additional 300 pairs with additional source text provided, resulting in 600 evaluation items. Additionally, in order to assess whether annotators assigned consistent ratings to the same S2ST pair across different intervals and thereby estimate annotation reliability, five randomly selected repeated items are also inserted throughout each survey. Consequently, each survey contains 630 evaluation items.

Specifically, inter-annotator variance measures the degree to which human evaluators differ in their ratings of the same item. In this evaluation, each audio sample was rated independently by multiple annotators, and disagreement is quantified as the standard deviation of scores across annotators for each item. A standard deviation of 0 indicates perfect agreement, while higher values reflect greater divergence in judgment.

A high standard deviation was observed in the absolute human evaluation scores, which can be attributed to the highly granular nature of the 0 – 100 Likert scale, as it naturally introduces subjective calibration differences across annotators. Although annotators consistently agreed on the relative ranking of the systems (i.e., AudioLLMs outperform cascaded systems), their individual criteria for which constitute an “acceptable” absolute score varied significantly. Despite the variance in absolute scoring, the aggregate trends robustly support our comparative conclusions.

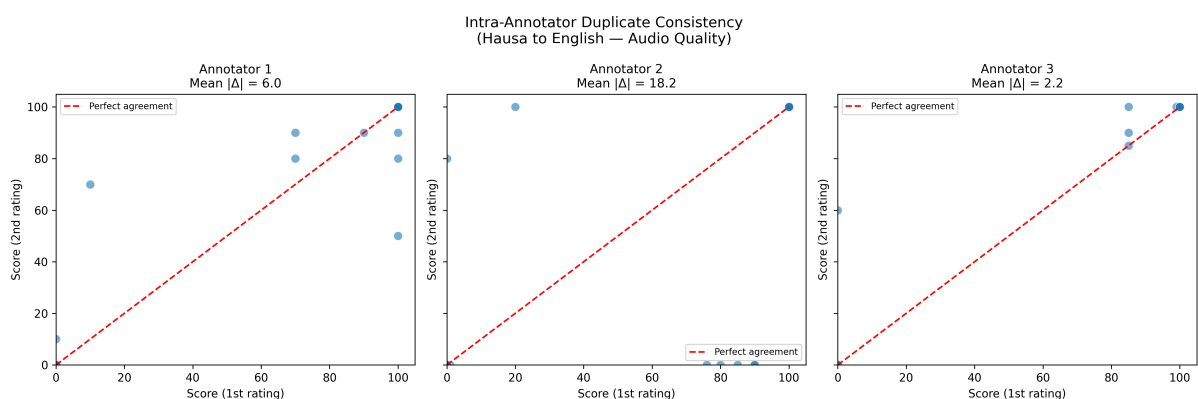


Figure 9: Intra-Annotator Consistency for S2ST translation quality for Hausa→English human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.

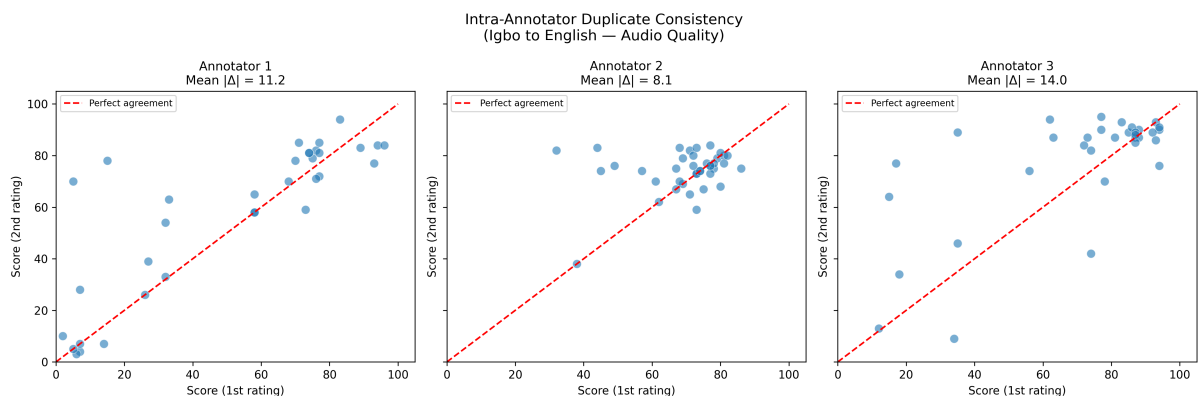


Figure 10: Intra-Annotator Consistency for S2ST translation quality for Igbo→English human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.

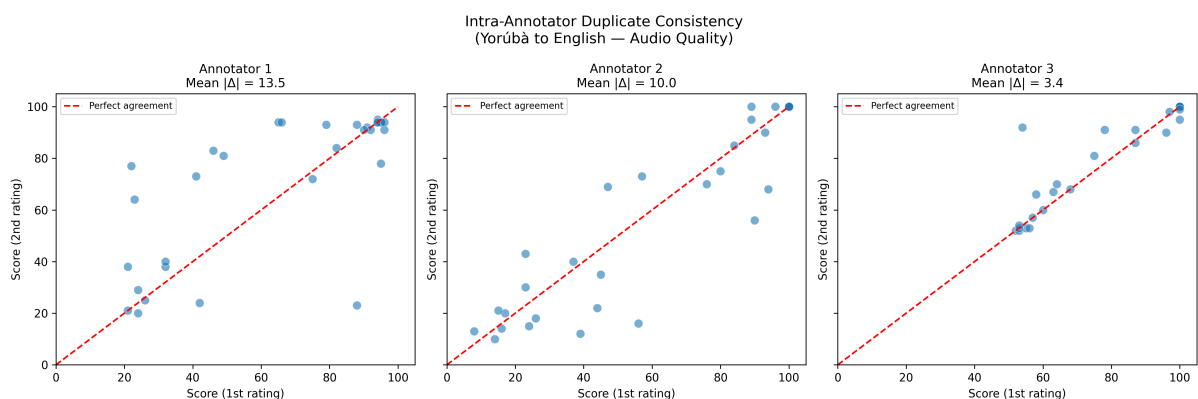


Figure 11: Intra-Annotator Consistency for S2ST translation quality for Yorùbá→English human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.

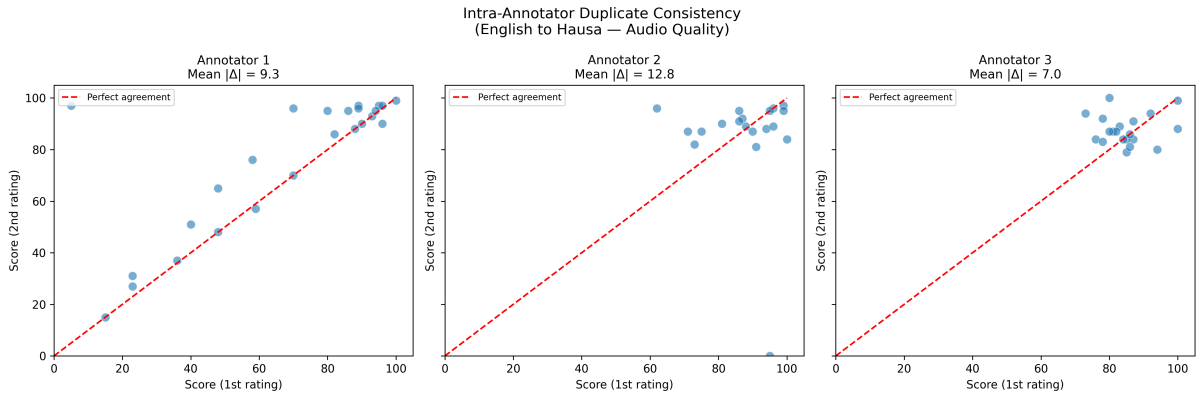


Figure 12: Intra-Annotator Consistency for S2ST translation quality for English→English human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.

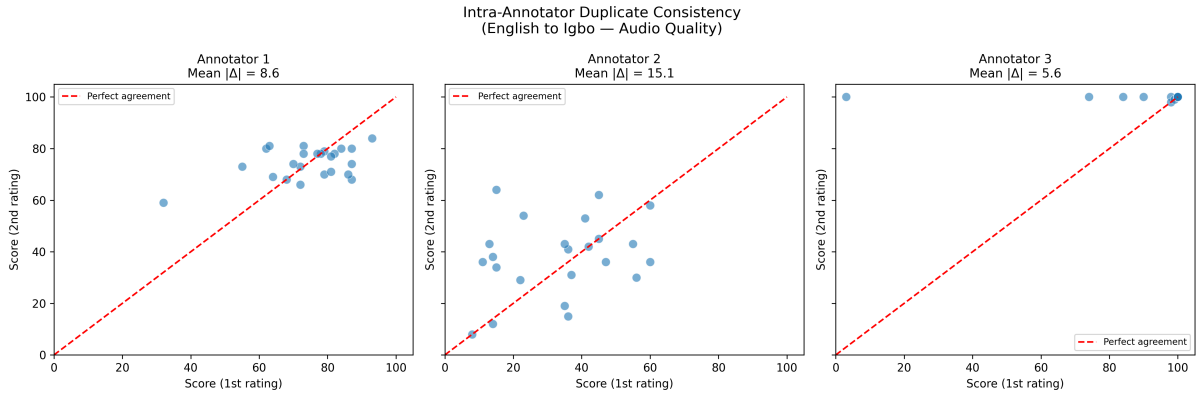


Figure 13: Intra-Annotator Consistency for S2ST translation quality for English→Igbo human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.

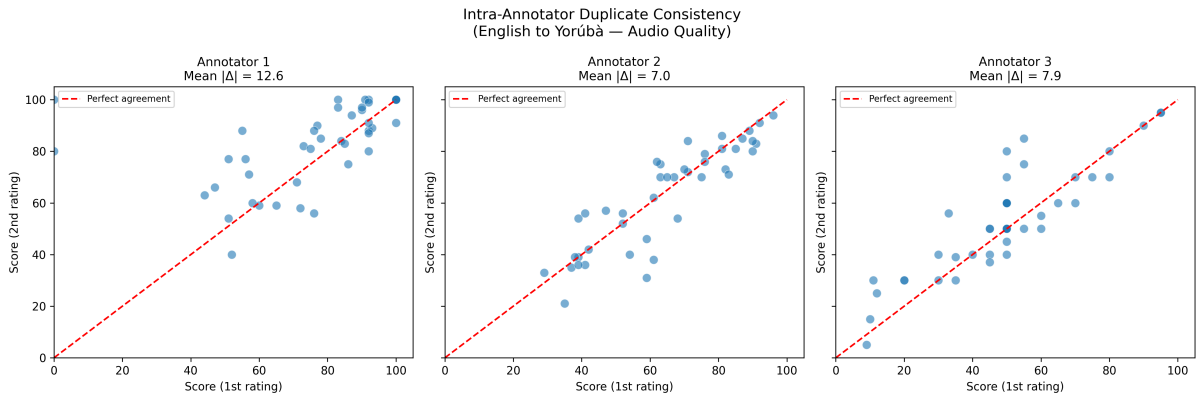
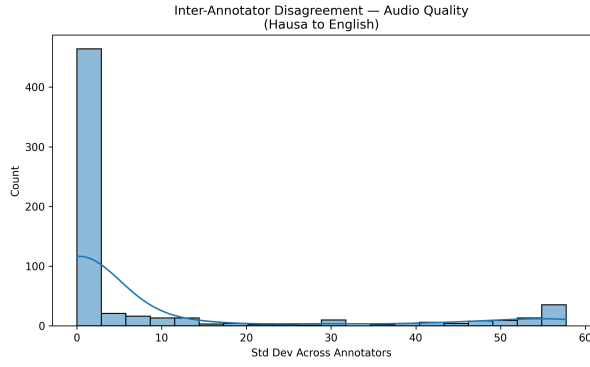
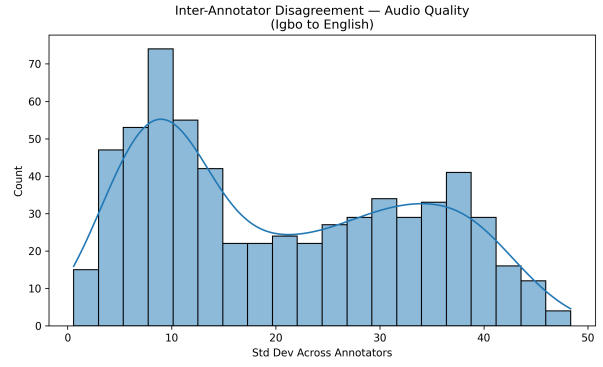


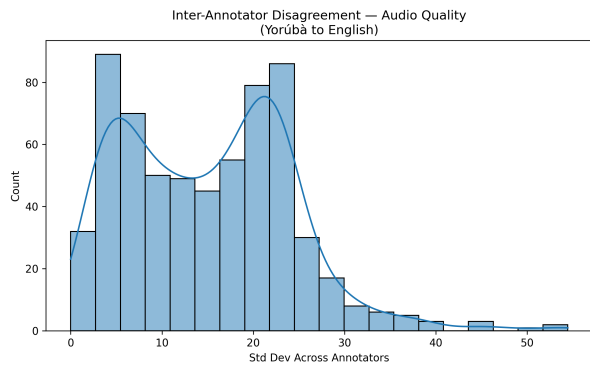
Figure 14: Intra-Annotator Consistency for S2ST translation quality for English→Yorùbá human evaluation. Points closer to the diagonal indicate higher self-consistency across different rating intervals.



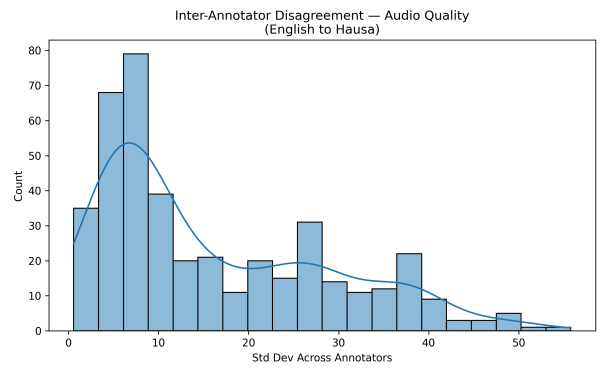
(a) Hausa→English



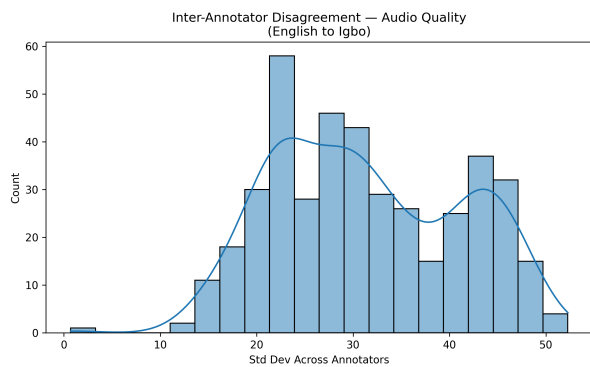
(b) Igbo→English



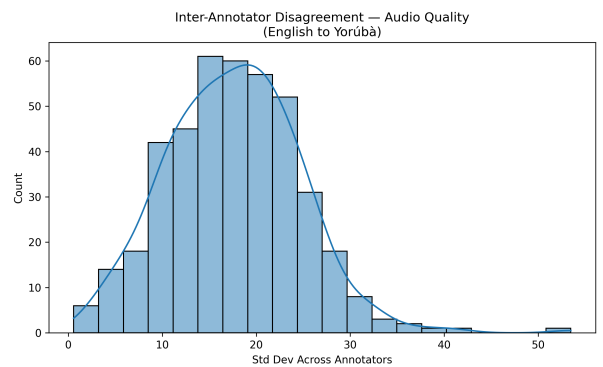
(c) Yorùbá →English



(d) English→Hausa



(e) English→Igbo



(f) English → Yorùbá

Figure 15: Inter-Annotator Variance for S2ST translation quality human evaluation across all language pairs.