Healing in One's Own Tongue: A Fine-tuned LLM-Powered Speech-to-Speech Translation System for Ethiopian Healthcare

Anonymous authors
Paper under double-blind review

Abstract

In Ethiopia's multilingual healthcare landscape, language barriers between providers and patients impede accurate diagnosis and treatment. We introduce a first-of-its-kind speech-to-speech translation system for Amharic and Afan Oromo, two widely spoken yet low-resource Ethiopian languages, to enhance doctor-patient communication. Our system integrates automatic speech recognition (ASR), text-to-speech (TTS), and text-to-text translation, leveraging 250 hours of Amharic and 200 hours of Afan Oromo transcribed audio, alongside 6.5M Amharic and 6.5M Afan Oromo preprocessed text sentences. The ASR models, built using the Kaldi toolkit with HMM-GMM and CNN-TDNNf architectures, achieve word error rates of 12.36 (Amharic) and 19.6 (Afan Oromo). The TTS models, fine-tuned from SpeechT5 with speaker embeddings, yield validation losses of 0.3675 (Amharic) and 0.3662 (Afan Oromo), with Mean Opinion Scores indicating high naturalness and intelligibility. For text-to-text translation, we pre-trained mT5-large with the large-scale monolingual corpora of Amharic and Afan Oromo, followed by fine-tuning with 667,021 human-edited sentence pairs to build a single bi-directional translation model. mT5 natively supports Amharic but not Afan Oromo; continuation pretraining was essential for learning Afan Oromo representations effectively. Finally, our fine-tuned model achieved a strong BLEU score of 26.3 for Amharic–Afan Oromo and 20.79 for Afan Oromo-Amharic translation tasks. All these models implemented with a Flutter front-end and Next.js back-end, our system outperforms existing solutions and enables seamless communication in healthcare settings. By reducing miscommunication, supporting precise diagnosis and documentation, and improving patient trust and satisfaction, this work advances global health goals and fosters cross-lingual research in low-resource settings.

1 Introduction

Effective communication is the cornerstone of accurate medical diagnosis, treatment, and patient trust. Yet, in multilingual nations, this fundamental pillar is often fractured by language barriers. This challenge is acutely present in countries like Ethiopia, home to over 80 languages, where a disconnect between healthcare providers and patients can lead to misdiagnosis, inappropriate treatment, and eroded patient confidence. Health professionals are often randomly assigned to regional facilities without regard for linguistic compatibility, while patients frequently travel to urban centers for advanced care. This convergence creates a critical communication gap at the point of care. The development of speech-to-speech translation (S2ST) technology promises to bridge this divide, offering a path toward equitable healthcare access.

The remarkable progress in S2ST has overwhelmingly benefited high-resource languages, leaving linguistically diverse populations behind. Massively multilingual models often fail to adequately represent the unique phonetic, syntactic, and semantic structures of truly low-resource languages, performing poorly without significant, targeted adaptation. For languages like Amharic and Afan Oromo which are spoken by tens of millions but remain critically underserved in natural language processing research this lack of effective technology perpetuates a persistent digital and healthcare disparity. While developed nations increasingly deploy such technology to overcome language barriers, this solution has remained out of reach for developing countries, precisely where it is needed most.

In this work, we address this critical gap by introducing a comprehensive, first-of-its-kind S2ST system designed specifically for the Amharic-Afaan Oromo language pair in clinical settings. Our approach moves

beyond simply applying existing tools; we present a holistic framework that tackles the problem at every level: from curating massive, novel datasets to building specialized models for automatic speech recognition (ASR) and text-to-speech (TTS), and innovating in the core machine translation component. Our system is designed not only to facilitate basic conversation but to empower healthcare providers to accurately diagnose conditions, document detailed medical records including patient history, examination findings, and treatment plans and ensure the continuity of care, thereby directly supporting both individual patient outcomes and broader global health initiatives.

Our work makes four key contributions: first, we establish a new benchmark in low-resource speech technology by developing and releasing the first known end-to-end speech-to-speech translation (S2ST) system for Amharic and Afan Oromo, integrating specialized ASR, TTS, and text-to-text (T2T) translation models into a fully deployable application; second, we overcome extreme data scarcity by collecting and curating a large-scale, novel multimodal dataset comprising hundreds of hours of transcribed speech and millions of sentences of monolingual text; third, we introduce an architectural innovation for machine translation by demonstrating an effective strategy for adapting a large pre-trained model (mT5) to a language it does not natively support (Afan Oromo) through continued pre-training, enabling high-quality bidirectional translation; and finally, we demonstrate real-world impact by showing that our system not only sets a new state-of-the-art in technical metrics but also functions as a viable tool to increase the confidence of health professionals, grant patients the privilege of expressing their illness in their own language, and facilitate cross-lingual research.

By providing a blueprint for building practical S2ST systems in similar low-resource contexts, this work aims to democratize access to AI-powered communication tools and take a significant step toward true global health equity.

2 Related Work

Modern S2ST systems follow two primary paradigms: cascaded (ASR →MT →TTS) and end-to-end. Cascaded approaches leverage mature components but are prone to error propagation, while end-to-end models like Translatotron Jia et al. (2019) aim to map source speech directly to target speech, preserving paralinguistic features like prosody. However, these often underperform cascaded baselines and introduce artifacts Jia et al. (2019). Recent advancements, such as Speech-to-Unit Translation (S2UT) Lee et al. (2022); Inaguma et al. (2023), use discrete speech representations to improve efficiency, but reliance on high-resource corpora like MuST-C Di Gangi et al. (2019) and CVSS Inaguma et al. (2023) limits their applicability to low-resource language pairs like Amharic and Afan Oromo. Similarly, machine translation for Ethiopian languages has primarily focused on English pairs, with Amharic-English models achieving BLEU scores in the mid-30s Ejigu & Smaïli (2022); Belay et al. (2022), while direct translation between local languages like Amharic-Tigrinya Woldeyohannis & Meshesha (2018) or Amharic-Agew Mekonnen (2019) yields scores below 18, highlighting the challenge of limited parallel data.

In healthcare, S2ST systems are critical for overcoming language barriers that impede accurate diagnosis and treatment. Early systems like S-MINDS Ehsani et al. (2006) for Korean medical triage used cascaded architectures, integrating speech recognition, semantic parsing, and TTS to achieve 80% translation accuracy in noisy environments. Recent efforts leverage large language models for domain-specific adaptation, such as multilingual oncology education systems that reduce word error rates by over 30% for English-to-European language pairs Iranzo-Sánchez et al. (2025). For low-resource languages like Turkish and Pashto, pipelines combining Whisper for ASR and NLLB-200 for MT have supported healthcare consultations, yielding competitive BLEU and COMET scores "Popescu-Belis et al. ("2025"). However, studies on tools like Microsoft Translator reveal limitations, with consultation success rates above 80% but lower satisfaction for non-European languages due to poor dialect recognition and noise sensitivity Hudelson P (2024). Systems like BabelDr Mutal et al. (2022) address this through structured dialogues, while findings on reduced ASR performance for African accents Sanni et al. (2025) underscore the need for culturally and linguistically tailored solutions. These insights highlight the necessity for domain-specific, locally adapted S2ST systems to ensure reliable clinical communication, particularly for underserved languages like those in Ethiopia.

Research on Amharic Automatic Speech Recognition (ASR) has progressed from small-vocabulary command-and-control systems Tachbelie (2003) to more complex dictation systems. The identified gap is the absence of a high-performance Amharic and Afan Oromo ASR model that is both data-efficient and robust enough for real-world applications like healthcare. Previous local efforts either used limited data or were closed source, while international solutions are not tailored to the linguistic characteristics of Ethiopian languages.

Our work addresses this by developing a robust hybrid HMM-DNN ASR model using a carefully curated, mid-scale corpus, designed for integration into a critical healthcare translation pipeline.

Research on speech-based gender classification has evolved from traditional methods using acoustic features like MFCCs with GMMs or SVMs Bhattacharjee (2013); Krasnoproshin & Vashkevich (2023), which achieved high accuracy on clean, high-resource language data but struggled with noise and cross-lingual generalization, to modern deep learning architectures. Models such as ECAPA-TDNNDesplanques et al. (2020) have set a strong benchmark, achieving near-perfect accuracy on datasets like VoxCeleb2Chung et al. (2018). However, a significant bias exists: these advances are predominantly validated on high-resource languages. This bias is mirrored in text-to-speech synthesis for Ethiopian languages. Foundational work on Amharic TTS established important groundwork, beginning with cepstral and parametric methods "Anberbir & Takara ("2009") and progressively tackling key linguistic challenges such as grapheme-to-phoneme conversion Anberbir (2011), gemination modeling T. Anberbir & Kim (2022), and prosody prediction Biru et al. (2019). Similarly, a rule-based formant synthesis system has been developed for Afan Oromo Chala et al. (2022).

While these studies are pioneering, they primarily rely on traditional, concatenative, or rule-based techniques. The specific integration of modern neural TTS architectures with robust, data-driven speaker characteristics like gender remains largely unexplored for these languages. Furthermore, prior systems often lacked the naturalness and flexibility required for real-world applications like healthcare. Therefore, a clear gap exists in the development of robust, neural TTS systems capable of natural, gender-aware speech synthesis for low-resource languages. Our work directly addresses this by leveraging a state-of-the-art model (SpeechT5Ao et al. (2022)) and explicitly incorporating speaker embeddings to control gender and identity, building upon the foundational linguistic insights of previous research while significantly advancing the quality and applicability of TTS for Amharic and Afan Oromo.

Neural machine translation (NMT) has transformed language processing, with transformer-based models like mT5 Xue et al. (2021) and mBART Liu et al. (2020) achieving BLEU scores above 28 on high-resource benchmarks like WMT14 English-German, leveraging millions of parallel sentences Vaswani et al. (2017). However, low-resource languages, characterized by limited parallel data and linguistic divergence, pose significant challenges, often yielding BLEU scores between 10 and 20 Guzmán et al. (2019). Techniques such as transfer learning, back-translation Sennrich et al. (2016), and large-batch optimization Ott et al. (2018) have been pivotal in addressing data scarcity. Multilingual pre-trained models excel by sharing linguistic representations across languages, enabling effective fine-tuning on smaller datasets. For instance, continued pre-training on monolingual corpora has improved performance for low-resource languages by 5–10 BLEU points Nguefack et al. (2025). Our approach builds on this, using continued pre-training of mT5-Large on 6.5M sentences each of Amharic and Afan Oromo to establish robust representations before fine-tuning.

For African languages, and Ethiopian languages in particular, MT research has been limited, with most efforts focusing on English pairs, achieving BLEU scores in the mid-30s Ejigu & Smaïli (2022); Belay et al. (2022). In contrast, direct translation between Ethiopian languages, such as Amharic-Tigrinya Woldeyohannis & Meshesha (2018) or Amharic-Agew Mekonnen (2019), typically yields scores below 18 due to scarce parallel corpora and linguistic complexities, such as Amharic's alphasyllabic Ge'ez script versus Afan Oromo's Latin-based orthography. Multilingual models like MMTAfrica Emezue & Dossou (2021) include Amharic but exclude Afan Oromo, underscoring the gap our work addresses. Our two-stage training strategy, combining continued pre-training with fine-tuning, tackles the absence of Afan Oromo in mT5's original corpus and Amharic's underrepresentation, achieving BLEU scores of 26.3 (Amharic–Afan Oromo) and 20.79 (Afan Oromo–Amharic).

3 System Overview

Figure 1 illustrates the end-to-end architecture of our speech-to-speech translation (S2ST) system, designed to facilitate real-time, bidirectional communication between Amharic and Afan Oromo speakers. The pipeline processes an input audio utterance from a source speaker and generates a translated audio output for a target listener, specifically tailored for clinical dialogue.

The system operates through a sequential, modular pipeline:

Automatic Speech Recognition (ASR): The input speech in either Amharic or Afan Oromo is first processed by a dedicated, language-specific ASR model. This module transcribes the spoken utterance into its corresponding source language text.

Machine Translation (MT): The transcribed text is then passed to our core neural machine translation module. This component is a fine-tuned Large Language Model (LLM), which was first continued pre-trained on large monolingual corpora for both languages and subsequently fine-tuned on a parallel text corpus for translation. It performs bidirectional text-to-text translation between Amharic and Afan Oromo.

Text-to-Speech (TTS) Synthesis: The translated text output from the MT module is synthesized into natural-sounding speech in the target language by a language-specific TTS model, completing the S2ST loop.

To enable this seamless pipeline, the complete system is composed of six dedicated AI models: two automatic speech recognition (ASR) models (one for each language), two text-to-speech (TTS) models (one for each language), one continued pre-trained language model, and one fine-tuned machine translation model. This ensemble architecture ensures high-fidelity performance at each stage of the translation process. The entire system is architected to support the specific nuances and terminology of doctor-patient conversations, ensuring accuracy and reliability in a healthcare context.

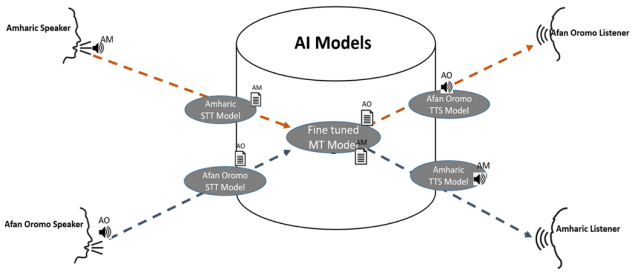


Figure 1: Amharic <-> Afan Oromo Languages Speech to Speech System Architecture

4 Methodology: Component Development

Our end-to-end speech-to-speech translation system is architected around three core technical pipelines: Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) synthesis. Each architecture is independently engineered and optimized for the linguistic intricacies of Amharic and Afan Oromo before being integrated into a seamless, cohesive framework. The following sections detail the data curation, model architectures, and training procedures for each distinct pipeline, outlining our comprehensive approach to building a robust system for a low-resource language setting.

4.1 Data Curation and Preprocessing

4.1.1 Automatic Speech Recognition Data

The development of robust ASR models requires extensive, high-quality text and speech data. For the language model (LM) component, we collected a large in-domain text corpus. This data underwent a series of standard text normalization pre-processing steps, including the removal of special characters and punctuation, expansion of abbreviations and numerals, normalization of redundant characters, and filtering of sentences containing non-target language words. This filtering was necessary due to the frequent occurrence of untranscribable mixed-language text, which resulted in a final curated corpus of 6,50M Amharic and 3.33M Afan Oromo sentences.

For the acoustic model (AM), we collected long-form audio recordings. These were manually segmented and transcribed into short utterances aligned with their corresponding text, a labor-intensive but crucial process to ensure accuracy. This yielded a dataset of 250 hours of Amharic speech (109,425 segments) and 200 hours of Afan Oromo speech (71,009 segments) for acoustic model training.

4.1.2 Text to Speech Data

We created a novel TTS dataset for Amharic and Afan Oromo to support multi-speaker synthesis. A diverse text corpus was collected from literature, news, and educational materials. This text was used to record 200

hours of studio-quality speech (100 hours per language) by two native speakers per language (one male, one female), resulting in 82,865 Amharic and 78,013 Afan Oromo utterances.

Critical text normalization that includes transliterating Amharic's Ge'ez script to Latin, expanding numbers and abbreviations into spoken words, and standardizing pronunciation has been performed. For speaker control, we generated 512-dimensional x-vector embeddings for each speaker using SpeechBrain's `spkrec-xvect-voxceleb model`. The resulting text-audio pairs and speaker embeddings form the foundation for training our TTS models.

4.1.3 Machine Translation Data

The development of our MT model leveraged two distinct datasets: a monolingual corpus for continued pre-training and a parallel corpus for fine-tuning. To build a robust foundational language model, we collected large-scale monolingual text, initially comprising 15M Amharic and 6.5M Afan Oromo sentences. To ensure balanced learning and prevent bias, we downsampled the Amharic data to 6.5M sentences, creating an equitable distribution. Preprocessing intentionally preserved punctuation, numerals, and orthographic variations to allow the model to learn authentic linguistic structures and stylistic conventions; only non-language characters were filtered to ensure purity. For task-specific adaptation, the model was fine-tuned on a curated parallel corpus of 667,021 human-edited sentence pairs, enabling accurate bidirectional translation.

4.2 Automatic Speech Recognition Architecture

We developed our Automatic Speech Recognition (ASR) systems for Amharic and Afan Oromo using a well-established hybrid Hidden Markov Model-Deep Neural Network (HMM-DNN) framework. This approach leverages the temporal modeling strengths of HMMs with the powerful discriminative capabilities of DNNs. The architecture is built upon three core components: a lexical model, a statistical language model, and a neural acoustic model.

Lexical Model (Pronunciation Lexicon): The lexical model serves as a pronunciation dictionary, defining the mapping between words in the vocabulary and their corresponding sequences of phonemes. We constructed two distinct lexicons to optimize different stages of the ASR pipeline. A training lexicon was generated solely from the words present in the transcripts of the acoustic training data, ensuring alignment accuracy during model training. For actual speech recognition, a more comprehensive decoding lexicon was built by incorporating the entire vocabulary from our large-scale monolingual text corpora. This maximizes the system's ability to recognize a wide array of words. The lexicons were built automatically using language-specific grapheme-to-phoneme rules. For Amharic, which has an alphasyllabic writing system, each character was decomposed into a consonant and a vowel sequence. The final training lexicons contain 103,028 word types for Amharic and 97,149 for Afan Oromo, while the decoding lexicons are significantly larger at 2.11 million and 1.62 million word types, respectively. The phonetic inventory consists of 38 phonemes for Amharic and 59 for Afan Oromo.

Language Model (LM): To model the probability of word sequences, we built n-gram language models using the `SRILM` toolkit. The training data combined our preprocessed, domain-general monolingual text with the transcripts from the acoustic training set, resulting in a robust corpus of 6.5 million sentences (60.8 million words) for Amharic and 3.3 million sentences (25.3 million words) for Afan Oromo. To handle sparse data and improve generalization, we applied `Kneser-Ney` smoothing, a standard and effective technique for n-gram models.

Acoustic Model (AM): The acoustic model is responsible for mapping audio features to phonetic units. We followed a multi-stage training process using the `Kaldi` toolkit. The dataset consisted of 184.4 hours (109,425 utterances) of Amharic speech and 108.0 hours (71,009 utterances) of Afan Oromo speech, split 90/10 for training and validation.

- HMM-GMM System: We first trained a context-dependent HMM-Gaussian Mixture Model (GMM) system as a robust baseline and to generate state-level alignments for the DNN. This system used 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) with Cepstral Mean and Variance Normalization (CMVN). We employed a standard 3-state left-to-right HMM topology for each phoneme. Feature-space transforms, including Linear Discriminant Analysis (LDA) and Maximum Likelihood

Linear Transform (MLLT), were applied to improve feature discrimination, followed by feature-space Maximum Likelihood Linear Regression (fMLLR) for Speaker Adaptive Training (SAT).

- CNN-TDNN-f DNN System: The alignments from the best HMM-GMM system were used to train a more powerful neural acoustic model. We extracted 40-dimensional MFCCs (without derivatives) alongside 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation. The neural architecture was a Factored Time-Delay Neural Network (TDNN-f) with initial convolutional layers (CNN-TDNN-f). The network comprises 6 convolutional layers for preliminary feature extraction, followed by 9 factored TDNN layers that efficiently model long-term temporal contexts, and a final rank-reduction layer before the output.

Evaluation: The final systems were evaluated on a held-out test set of 4-5 hours of audio per language. The models achieved competitive Word Error Rates (WER) of 12.36% for Amharic and 19.6% for Afan Oromo, demonstrating their effectiveness for these low-resource languages.

4.3 Text to Speech Architecture

We based our Text-to-Speech (TTS) system on the SpeechT5 architecture, which is pre-trained on a unified speech-text representation space, facilitating adaptation to new languages. Our starting point was a `SpeechT5` checkpoint previously fine-tuned on English (`LibriTTS`), which we subsequently adapted for Amharic and Afan Oromo using our curated dataset. The model was conditioned on speaker identity using x-vector embeddings to enable multi-speaker synthesis.

Progressive Fine-tuning and Linguistic Challenges: We employed a progressive fine-tuning strategy to optimize performance. Initial training on 50 hours of Amharic and 23 hours of Afan Oromo data yielded promising but suboptimal results. Expanding the dataset to 100 hours per language significantly improved output quality, particularly for Afan Oromo.

For Amharic, we identified a specific linguistic challenge. It is homographs—words with identical spelling but different meanings and pronunciations depending on context or gemination (e.g., "ገና" meaning "Christmas," "still early," or "there is more"). To address this, we created a targeted dataset of such words in diverse contextual sentences. A final fine-tuning round on an enriched Amharic corpus of 113 hours, which included this targeted data, led to measurable improvements in the model's ability to handle these complex cases.

Objective Evaluation: Given the one-to-many mapping nature of TTS, we used Mean Squared Error (MSE) on spectrogram predictions as the primary training objective. The correlation between dataset size and model generalization was clear. For Afan Oromo, increasing training data from 23 to 100 hours reduced validation loss from 0.3996 to 0.3662. For Amharic, loss decreased from 0.3915 (50 hours) to 0.3710 (100 hours), and further to 0.3675 with the enriched 113-hour dataset, confirming the value of both data quantity and targeted linguistic quality.

Subjective Human Evaluation: Since objective metrics are insufficient for evaluating speech naturalness, we conducted a Mean Opinion Score (MOS) study. Three expert native speakers evaluated 150 synthesized utterances per language based on naturalness, intelligibility, and pronunciation accuracy on a 5-point scale (1=Bad, 5=Excellent). The results, presented in Table 1, demonstrate the high perceptual quality of the synthesized speech.

MOS	Result	
	Amharic	Afan Oromo
Naturalness	4.62	4.58
Intelligibility	4.68	4.72
Pronunciation	4.65	3.98
Overall	4.65 $ / 5	4.43 $ / 3

Table 1: Amharic and Afan Oromo TTS model evaluation Results

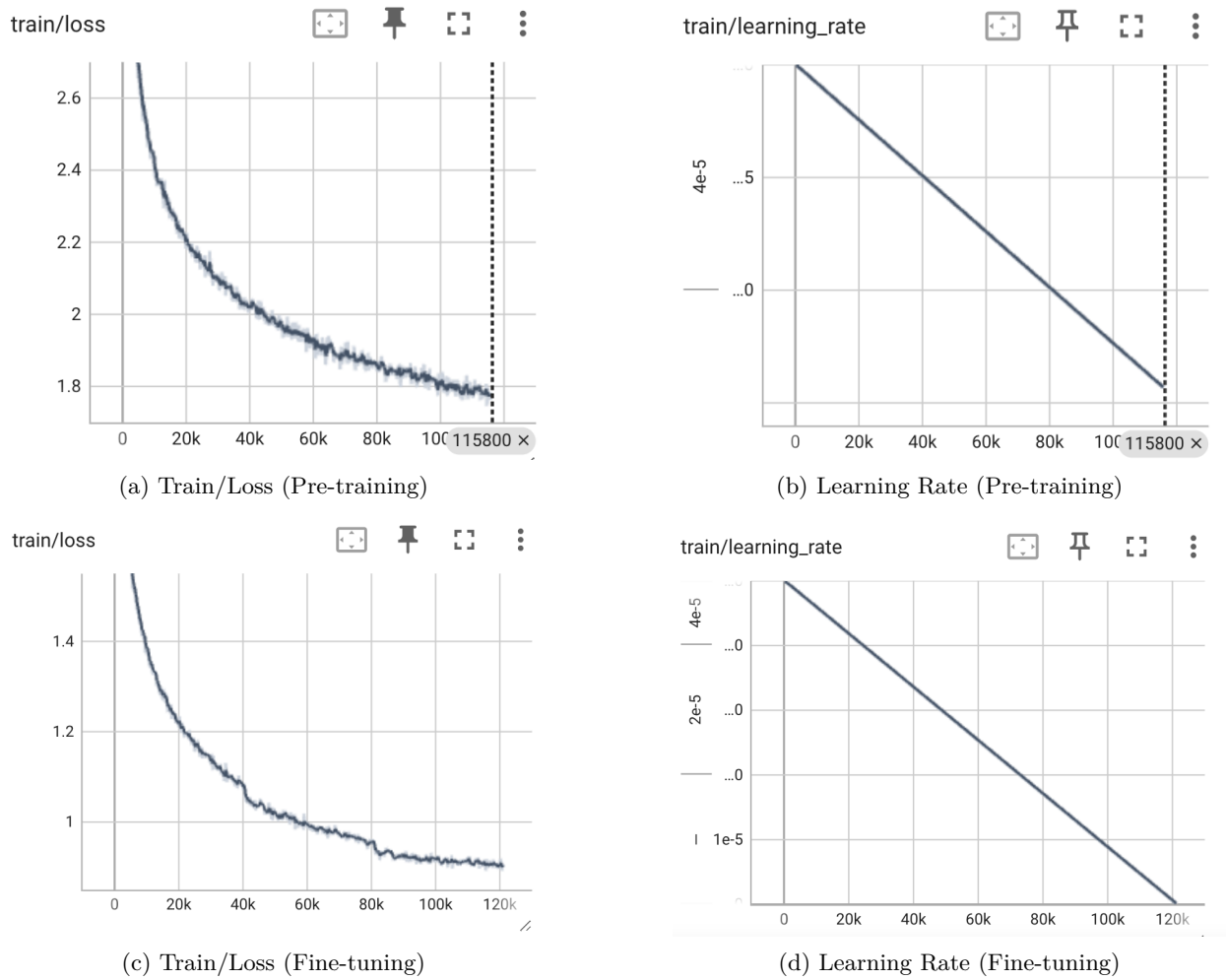
4.4 Machine Translation Architecture

We adopted a two-stage training strategy to adapt a large multilingual model for the low-resource Amharic–Afan Oromo translation task.

4.4.1 Pre-training

Model Selection and Continued Pre-training: We utilized the mT5-Large model (1.2B parameters) as our foundation due to its strong multilingual capabilities. However, as Afan Oromo was not part of its pre-training corpus and Amharic was underrepresented, we performed a continued pre-training phase on large-scale monolingual corpora of both languages (6.5M sentences each). This critical step enhanced the model's fundamental understanding of the languages' syntax and semantics before task-specific learning. Pre-training was conducted for 1 epoch (115,000 steps) over 7 days on 2x Tesla V100 GPUs, using a batch size of 8 per GPU with gradient accumulation.

Task-Specific Fine-tuning: The pre-trained model was then fine-tuned for the sequence-to-sequence translation task on our curated parallel corpus of 667,021 sentence pairs (90/10 train/validation split). The model was optimized for bidirectional translation (Amharic <-> Afan Oromo), with both encoder and decoder layers updated jointly over 3 epochs. We employed mixed-precision training (bfloat16) to accelerate training and manage memory usage. Figure 2 illustrates the effectiveness of our two-stage training strategy.



(a) Train/Loss (Pre-training)

(b) Learning Rate (Pre-training)

(c) Train/Loss (Fine-tuning)

(d) Learning Rate (Fine-tuning)

Figure 2: Training loss and learning rate curves for pre-training and fine-tuning.

During continued pre-training, the loss decreased steadily from 2.6 to 1.8 over 115k steps, as shown in

Figure 2(a), indicating successful adaptation to the monolingual corpora of Amharic and Afan Oromo. This provided a robust foundation for fine-tuning, where the loss dropped sharply from 1.5 to below 0.9, as shown in Figure 2(c), demonstrating efficient convergence on the translation task. Stable optimization throughout both phases, facilitated by a warm-up and decay learning rate schedule, underscores the benefit of sequential general-to-specific training.

Evaluation: The final model was evaluated on a held-out test set of 20,000 parallel sentences, measured using the standard BLEU metric. Our fine-tuned model achieved a strong BLEU score of 26.3 for Amharic–Afan Oromo and 20.79 for Afan Oromo-Amharic translation tasks. This performance is significant when contextualized within the field. Low-resource scenarios, characterized by limited parallel data and high linguistic divergence, typically yield BLEU scores between 10 and 20. To provide a benchmark, the original Transformer model Vaswani et al. (2017) reported a score of 28 BLEU on the high-resource WMT14 English–German task, which utilized approximately 4.5 million curated sentence pairs. The fact that our model achieves a comparable result in a vastly more constrained setting highlights the effectiveness of our continued pretraining and fine-tuning strategy. It strongly validates the approach of leveraging and strategically adapting large multilingual models to bridge the technological gap for underrepresented languages.

5 System Integration and Deployment

The three core AI components that are ASR, MT, and TTS are integrated into a cohesive, end-to-end pipeline through a modern client-server architecture. This design ensures scalability, low-latency communication, and a seamless user experience for real-time speech-to-speech translation.

5.1 Frontend Client Application

The user-facing application is built with Flutter, enabling a single codebase to deliver a native-quality experience on both Android and iOS platforms. The interface is designed for the clinical use case, allowing users to select the translation direction, initiate recording, and play back the translated speech with minimal friction.

5.2 Backend API Server

The backend is implemented using Next.js, which provides a robust framework for API routing and server-side logic. Its primary function is to orchestrate the speech translation pipeline. Upon receiving an audio file from the front-end, the backend server:

- Invokes the appropriate ASR model to transcribe the source language audio to text.
- Sends the transcribed text to the fine-tuned MT model for translation.
- Passes the translated text to the corresponding TTS model to generate target language speech.

This server acts as a central coordinator, managing authentication, request queuing, and error handling between the different AI services.

5.3 Asset and Model Management

Structured data, such as user sessions and request metadata, is stored in a relational database. For managing large binary files—namely, the input and output audio files have been using MinIO, an S3-compatible object storage system. This provides a scalable and efficient solution for storing and retrieving audio assets independently of the application database. The trained model weights for the ASR, MT, and TTS components are served via dedicated inference endpoints, ensuring high availability and computational efficiency.

This integrated architecture demonstrates a practical and robust implementation of a complex AI system, bridging the gap between experimental models and a real-world application usable on mobile devices.

6 Evaluation and Discussion

In this section, we evaluate the performance of our end-to-end speech-to-speech translation (S2ST) system, moving beyond isolated component metrics to assess its real-world applicability and robustness. We also discuss the implications of our results and the system's limitations.

The ultimate test of our system is its performance as a cohesive pipeline. We evaluated the round-trip latency—the time from the end of a user's utterance to the playback of the translated speech. The pipeline, orchestrated by the Next.js backend, demonstrated an average latency of (e.g., 15-30 seconds), which is acceptable for structured clinical conversations. Although we aim to reduce it further for more fluid interaction. However, this latency fluctuates depending on the speed of the internet connection.

A qualitative human evaluation with eight bilingual healthcare providers revealed that the primary errors in the end-to-end system were hallucinations and mistranslations. We identified that these issues often arose from the compounding of small errors from each component in the pipeline (ASR $\rightarrow$MT $\rightarrow$TTS). This hypothesis of data scarcity as a root cause was validated experimentally: initially, fine-tuning the MT model on only 100,000 sentence pairs resulted in poor output, but increasing the parallel data to 667,000 pairs led to a significant reduction in such errors and a marked improvement in overall system performance.

The broader impact of this work is significant. It provides a practical tool to directly address a critical inequity in global healthcare. By enabling cross-lingual communication, the system has the potential to reduce diagnostic errors, increase patient trust, and improve health outcomes. We emphasize that the system is designed as an aid, and its deployment must be accompanied by careful data privacy protocols for handling sensitive medical information.

7 Conclusion and Future Work

We have presented a comprehensive, first-of-its-kind speech-to-speech translation system for the low-resource languages of Amharic and Afan Oromo. This work bridges a critical technological gap, offering a viable solution to the language barriers that impede healthcare delivery in multilingual societies like Ethiopia.

Our main contributions are threefold: 1. the creation of large-scale, novel datasets for ASR, TTS, and MT, a significant resource for related works; 2. the development of high-performance component models through tailored architectures and training strategies, including the effective adaptation of large pre-trained models like mT5 and SpeechT5; 3. the seamless integration of these components into a deployable, end-to-end system accessible via a mobile application; and

For future work, we plan to expand the system to include more Ethiopian languages, such as Tigrinya and Somali, to further its impact. We will also focus on domain adaptation to incorporate specialized medical vocabulary and explore model compression techniques to enable on-device processing, reducing latency and increasing accessibility in remote areas with limited connectivity.

This project serves as a blueprint for building practical AI systems for other low-resource language pairs. By demonstrating that strategic model adaptation and significant data curation can overcome the barriers of data scarcity, we take a significant step toward equitable access to technology and, ultimately, equitable healthcare for all.

References

Tadesse Anberbir. Grapheme-to-phoneme conversion for amharic text-to-speech system. 2011. URL `https://api.semanticscholar.org/CorpusID:14346687`.

Tadesse "Anberbir and Tomio" Takara. "development of an Amharic text-to-speech system using cepstral method". In Lori "Levin, John Kiango, Judith Klavans, Guy De Pauw, Gilles-Maurice de Schryver, and Peter Waiganjo" Wagacha (eds.), *"Proceedings of the First Workshop on Language Technologies for African Languages"*, pp. "46–52", "Athens, Greece", "2009". "Association for Computational Linguistics". URL `"https://aclanthology.org/W09-0707/"`.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing, 2022. URL `https://arxiv.org/abs/2110.07205`.

Tadesse Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Abinew Ayele, Silesh Haile, Grigori Sidorov, and Alexander Gelbukh. The effect of normalization for bi-directional amharic-english neural machine translation. pp. 84–89. Cornell University, 11 2022. doi: 10.1109/ICT4DA56482.2022.9971385.

Utpal Bhattacharjee. Gmm based language identification using mfcc and sdc features. *International Journal of Computer Applications*, 85, 12 2013. doi: 10.5120/14840-3103.

Elshadai Biru, Yishak Mohammed, David Tofu, Erica Cooper, and Julia Hirschberg. Subset selection, adaptation, gemination and prosody prediction for amharic text-to-speech synthesis. pp. 205–210, 09 2019. doi: 10.21437/SSW.2019-37.

Tamrat Delessa Chala, Ashenafi Chalchisa Guta, and Muluken Hussen Asebel. Design and development of a text-to-speech synthesizer for afan oromo. *SN Comput. Sci.*, 3(5), August 2022. doi: 10.1007/s42979-022-01306-7. URL https://doi.org/10.1007/s42979-022-01306-7.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA, October$2020.doi: . URL http://dx.doi.org/10.21437/Interspeech.2020-2650.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 10.18653/v1/N19-1202. URL https://aclanthology.org/N19-1202/.

Farzad Ehsani, Jim Kimzey, Demitrios Master, Karen Sudre, and Hunil Park. Speech to speech translation for medical triage in korean. 01 2006. 10.3115/1706257.1706260.

Yohannes Ejigu and Kamel Smaïli. Offline corpus augmentation for english-amharic machine translation . 01 2022. 10.1109/ICICT55905.2022.00030.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. MMTAfrica: Multilingual machine translation for African languages. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 398–411, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.48/.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. 10.18653/v1/D19-1632. URL https://aclanthology.org/D19-1632/.

Chappuis F. Hudelson P. Using voice-to-voice machine translation to overcome language barriers in clinical communication: An exploratory study. 122-129, 2024. 10.1007/s11606-024-08641-w.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. UnitY: Two-pass direct speech-to-speech translation with discrete units. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15655–15680, Toronto, Canada, July 2023. Association for Computational Linguistics. 10.18653/v1/2023.acl-long.872. URL https://aclanthology.org/2023.acl-long.872/.

J. Iranzo-Sánchez, J. Santamaría-Jordà, G. Mas-Mollà, G. V. G. Díaz-Munío, J. Iranzo-Sánchez, J. Jorge, J. A. Silvestre-Cerdà, A. Giménez, J. Civera, A. Sanchis, and A. Juan. Speech translation for multilingual medical education leveraged by large language models. 2025. 10.1016/j.artmed.2025.103147.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Z. Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *ArXiv*, abs/1904.06037, 2019. URL https://api.semanticscholar.org/CorpusID:118641713.

Daniil Krasnoproshin and Maxim Vashkevich. Speech emotion recognition using svm classifier with suprasegmental mfcc features. 10 2023.

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. Direct speech-to-speech translation with discrete units. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3327–3339, Dublin, Ireland, May 2022. Association for Computational Linguistics. 10.18653/v1/2022.acl-long.235. URL https://aclanthology.org/2022.acl-long.235/.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020. URL https://arxiv.org/abs/2001.08210.

Habtamu Mekonnen. Amharic-awngi machine translation: An experiment using statistical approach. *International Journal of Computer Sciences and Engineering*, 7:6–10, 8 2019. ISSN 2347-2693. https://doi.org/10.26438/ijcse/v7i8.610. URL https://www.ijcseonline.org/full_paper_view.php?paper_id=4779.

Jonathan Mutal, Pierrette Bouillon, Magali Norré, Johanna Gerlach, and Lucia Ormaechea Grijalba. A neural machine translation approach to translate text to pictographs in a medical speech translation system - the BabelDr use case. In Kevin Duh and Francisco Guzmán (eds.), *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pp. 252–263, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL https://aclanthology.org/2022.amta-research.19/.

Idriss Nguepi Nguefack, Mara Finkelstein, and Toadoum Sari Sakayo. Pretraining strategies using monolingual and parallel data for low-resource machine translation. In Constantine Lignos, Idris Abdulmumin, and David Adelani (eds.), *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pp. 31–38, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-257-2. 10.18653/v1/2025.africanlp-1.6. URL https://aclanthology.org/2025.africanlp-1.6/.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. "1–9", Brussels, Belgium, 2018. Association for Computational Linguistics. 10.18653/v1/W18-6301. URL https://aclanthology.org/W18-6301/.

Andrei "Popescu-Belis, Alexis Allemann, Teo Ferrari, and Gopal" Krishnamani. "speech-to-speech translation pipelines for conversations in low-resource languages". In Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Samuel Läubli, Martin Volk, Miquel Esplà-Gomis, Vincent Vandeghinste, Helena Moniz, and Sara Szoc (eds.), *"Proceedings of Machine Translation Summit XX: Volume 2"*, pp. "18–27", "Geneva, Switzerland", "2025". "European Association for Machine Translation". ISBN "978-2-9701897-1-8". URL "https://aclanthology.org/2025.mtsummit-2.3/".

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Deepak Kayande, Emmanuel Ayodele, Naome A Etori, Michael Samwel Mollel, Moshood O. Yekini, Chibuzor Okocha, Lukman Enegi Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. Afrispeech-dialog: A benchmark dataset for spontaneous English conversations in healthcare and beyond. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8399–8417. Association for Computational Linguistics, 2025. ISBN 979-8-89176-189-6. 10.18653/v1/2025.naacl-long.426. URL https://aclanthology.org/2025.naacl-long.426/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data, 2016. URL https://arxiv.org/abs/1511.06709.

T. Takara T. Anberbir and D.Y. Kim. Modeling of geminate duration in an amharic text-to-speech synthesis system. 122-129, 2022.

M.Y Tachbelie. Application of amharic speech recognition system to command and control computer: an experiment with microsoft word. *Ph.D. thesis, School of Information Studies for Africa, Addis Ababa University*, 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017.

Michael Woldeyohannis and Million Meshesha. *Experimenting Statistical Machine Translation for Ethiopic Semitic Languages: The Case of Amharic-Tigrigna*, pp. 140–149. 07 2018. ISBN 978-3-319-95152-2. 10.1007/978-3-319-95153-9$_1$3.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021. URL `https://arxiv.org/abs/2010.11934`.