AURA-ST: Acoustic-Unconstrained Residual Architecture for Speech Translation

Barathi Ganesh HB[1], Michal Ptaszynski[1], Jairam R[2], Reshma Unnikrishnan[2]

[1]Kitami Institute of Technology, Kitami, Hokkaido 090-0015, Japan

[2]RBG AI Research, RBG.AI, SREC Incubation Center, Coimbatore, Tamil Nadu 641022, India

hbbg.jp@gmail.com, michal@mail.kitami-it.ac.jp, research@rbg.ai

Abstract

We present AURA-ST, a three-stage modular pipeline for low-resource speech-to-text translation submitted to the IWSLT 2026 African-Celtic Track 1. The architecture bypasses traditional cross-attention between audio and text modalities by treating projected acoustic representations as a native token prefix to a frozen large language model. A dual-stream encoder captures linguistic and paralinguistic features via a jointly trained semantic and a paralinguistic encoder. A convolutional subsampler then bridges the modality gap through a $4\times$ temporal compression and a linear projection into the LLM embedding space. Finally, a MLP-targeted Low-Rank Adaptation adapter fine-tunes the frozen Gemma-4-E2B backbone for translation without catastrophic forgetting of base language model knowledge. We further identify and resolve the incompatibility between standard PEFT attention-level adapter injection and the Gemma-4 Per-Layer Embedding architecture that tends to cause gradient isolation. Trained on the IWSLT 2026 Track 1 data covering Hausa, Igbo, and Yoruba, the final system achieves a best proxy teacher-forced SacreBLEU of **91.29** on the validation set at Phase 3, with Phase 1 speech encoder validation loss converging to **0.651**.

1 Introduction

End-to-end speech translation (ST) for low-resource African and Celtic languages presents extreme challenges like severe data scarcity, high tonal and dialectal variance, and the absence of large in-domain pretrained acoustic models. Prior work has largely addressed this through encoder-decoder architectures with cross-attention (Bérard et al., 2016; Weiss et al., 2017), or by cascading automatic speech recognition (ASR) with machine translation (MT) (Sperber and Paulik, 2020). While cascaded systems offer modularity, they propagate ASR errors into the translation stage. End-to-end systems avoid this but are data-hungry and difficult to adapt when source-side data is scarce.

More recent approaches leveraging pretrained models such as Whisper (Radford et al., 2023) and SeamlessM4T (Barrault et al., 2023) have demonstrated strong multilingual ST, but their heavy cross-attention modules impose significant memory overhead and limit adaptation flexibility on constrained hardware. In parallel, large language models (LLMs) have been connected to audio encoders via lightweight projection modules in systems such as Flamingo (Alayrac et al., 2022), LLaSA (Xie et al., 2024), and Qwen-Audio (Chu et al., 2023), demonstrating that a frozen LLM can serve as a powerful generation backbone for multimodal inputs when the projection layer is adequately trained. These observations motivate us to stitch a dual-stream acoustic encoder to a frozen causal LLM through a lightweight convolutional bridge, then adapt only the LLM's MLP layers via Low-Rank Adaptation (LoRA) (Hu et al., 2022) for translation. AURA-ST makes following concrete contributions:

- A dual-stream acoustic encoder that jointly models phonetic and paralinguistic information for tonal low-resource languages through a multi-objective training regime combining masked language modeling (MLM), contrastive learning, codebook diversity enforcement, and auxiliary classification.
- A convolutional subsampler trained to bridge the modality gap without cross-attention, using a prefix stitching strategy that natively embeds audio into the LLM's causal token stream.
- Identification of a structural incompatibility between PEFT attention-level LoRA and Gemma-4's Per-Layer Embedding (PLE) architecture and its resolution via MLP injection.
- Reproducible inference pipeline with pre-trained weights are made available at: `https://github.com/rbg-research/IWSLT-2026-AURA-ST`.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 247–254

2 Related Works

Self-supervised speech models such as wav2vec 2.0 (Baevski et al., 2020) and w2v-BERT (Chung et al., 2021) provide robust frame-level acoustic representations pretrained on massive multilingual corpora. The w2v-BERT architecture extends the MLM paradigm to speech, training both a contrastive objective and a BERT-style MLM objective on quantised latents simultaneously. Speaker-discriminative ResNet architectures trained on VoxCeleb (Zeinali et al., 2019) have shown strong cross-lingual generalisation as paralinguistic encoders and are commonly used for zero-shot speaker verification.

AURA-ST's dual-stream design draws inspiration from multi-view representation learning (Tian et al., 2020), where two complementary views of the same signal are contrasted to learn a richer shared representation than either view alone. The InfoNCE contrastive objective we employ between clean and augmented audio views directly follows van den Oord et al. (2018).

The convolutional subsampler used in our bridge module adapts the downsampling strategy of Li et al. (2021), which showed that a stack of 1D convolutions with stride 2 is sufficient to compress speech frame sequences to LLM-compatible lengths while preserving linguistically relevant content. Unlike its full fine-tuning approach, we train only the subsampler while keeping both the acoustic encoder and the LLM backbone fully frozen, significantly reducing the number of trainable parameters.

LoRA (Hu et al., 2022) has seen rapid adoption for efficient LLM fine-tuning by decomposing weight updates into low-rank matrices. Its application to speech-language models was explored in Fathullah et al. (2024), who showed that frozen LLMs can be adapted for ASR via instruction prompting combined with LoRA. Our work extends this line to the translation task and specifically addresses the architectural incompatibility that arises when LoRA is applied to LLMs with custom PLE-based positional encodings.

3 Data and Statistics

We use the data released for IWSLT 2026 Track 1 (Low-Resource African-Celtic Speech Translation), targeting Hausa (ha), Igbo (ig), and Yoruba (yo) as source languages with English (en) as the target (Maltais et al., 2026; Adelani et al., 2026).

Corpus Construction: All audio files are resampled to 16 kHz mono with a maximum duration per utterance is capped at 30 seconds (1788 files were eliminated). Short utterances are padded to a minimum of 1 second with zero-padding. The corpus covers **194 unique speakers**, **3 source languages**, and **2 annotated genders** derived from a pretrained model[1]. Speaker, language, and gender IDs are used as auxiliary supervision targets during Phase 1 encoder training.

Quality Filtering: For Phase 2 (modality alignment) and Phase 3 (translation), additional text-quality filters are applied. Entries with missing translation strings (734 utterances) or translations exceeding 1,500 characters are excluded, reducing the usable pool by roughly 1.6% relative to Phase 1.

Phase	Objective	Train	Val
Phase 1	Enc. pretraining	37,383	9,346
Phase 2	Modality alignment	36,797	9,198
Phase 3	S2TT fine-tuning	36,797	9,198

Table 1: Dataset split sizes per training phase after quality filtering.

Audio Augmentation: Robust speech representation learning under low-resource conditions requires strong augmentation. During Phase 1, 70% of training samples receive one or more of telephonic channel simulation, environment noise, RIR convolution, and music overlay augmentation, with SNR uniformly sampled from $[5, 20]$ dB and bit-depth sampled from $\{8, 16, 24, 32\}$ bits. The remaining 30% receive light resampling only. Validation data is never augmented.

Feature Extraction: Two parallel feature streams are extracted per utterance,

- **Semantic features:** Log-filterbank features at 16 kHz processed by the W2V-BERT feature extractor, producing $T \times 160$-dimensional input feature frames.
- **Acoustic features:** 80-dimensional log-Mel spectrograms computed with an STFT of 400-sample window (25 ms), 160-sample hop (10 ms), and 80 Mel bins, producing $T' \times 80$ frames.

Both streams are padded to the batch maximum length and accompanied by a binary attention mask indicating valid (non-padded) positions.

[1]Accessed on April, 2026. https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech

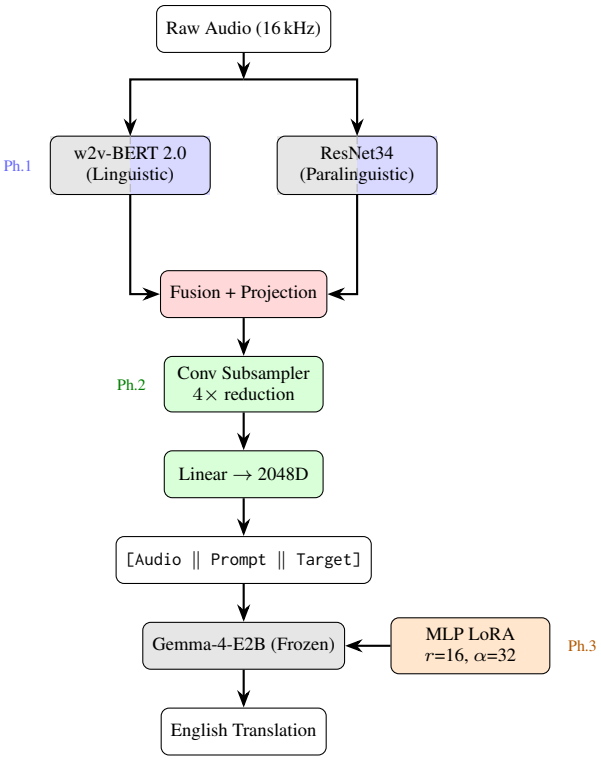

Figure 1: AURA-ST pipeline. Grey = frozen; Blue = partially trainable (Phase 1); Red = trained in Phase 1 (Fusion); Green = trained in Phase 2; Orange = LoRA (Phase 3).

4 System Architecture

AURA-ST is a fully modular three-phase pipeline. Each phase introduces a distinct component where earlier components are frozen in subsequent phases. Figure 1 provides a schematic overview.

4.1 Phase 1: Dual-Stream Acoustic Encoding

The foundational acoustic encoder ($\mathcal{E}$) fuses two parallel streams operating on the same audio signal.

Linguistic Stream: We employ partially trainable **w2v-BERT 2.0**[2] (Chung et al., 2021), a model pretrained on 4.5M hours of multilingual speech via a joint contrastive and masked speech modeling objective. It produces $T \times 1024$-dimensional frame-level representations ($\mathbf{H}^{\text{sem}}$) that encode rich phonetic, phonological, and sub-word transition information. The convolutional feature extractor inside w2v-BERT applies a stride-product of approximately 320, mapping 16 kHz audio to roughly 50 feature frames per second.

Paralinguistic Stream: A partially trainable **ResNet34** backbone[3] processes the 80-dimensional log-Mel spectrogram ($\mathbf{M} \in \mathbb{R}^{T' \times 80}$) and produces a global 256-dimensional embedding ($\mathbf{g} \in \mathbb{R}^{256}$) via average pooling over time. This stream was pretrained for speaker verification on VoxCeleb and generalises well to cross-lingual paralinguistic capture. For tonal languages (Yoruba: 3 tones; Igbo: 2 tones), this stream is expected to carry F0 and duration-modulated cues absent from the purely phonemic w2v-BERT representation.

Fusion: The global acoustic embedding ($\mathbf{g}$) is broadcast and concatenated with each frame of $\mathbf{H}^{\text{sem}}$ producing fused representations ($\tilde{\mathbf{H}} \in \mathbb{R}^{T \times d_f}$) that integrate both streams at every time step.

$$\tilde{\mathbf{H}}_t = \mathbf{W}_f \left[\mathbf{H}_t^{\text{sem}} \,\|\, \mathbf{g}\right] + \mathbf{b}_f \tag{1}$$

Multi-Task Training Objective. The encoder is optimised jointly with four loss components:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{cont}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + w_{\text{aux}}(s) \cdot \mathcal{L}_{\text{aux}} \tag{2}$$

MLM loss ($\mathcal{L}_{\text{MLM}}$): Random time spans are masked in the subsampled feature sequence and the model must predict the Gumbel-VQ-quantised target code for each masked frame. The loss is computed per sequence as an InfoNCE objective (van den Oord et al., 2018) over a set of distractors drawn from the same batch:

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{|M|} \sum_{m \in M} \log \frac{\exp(\hat{\mathbf{z}}_m \cdot \mathbf{q}_m / \kappa)}{\sum_j \exp(\hat{\mathbf{z}}_m \cdot \mathbf{q}_j / \kappa)} \tag{3}$$

where $\hat{\mathbf{z}}_m$ is the predicted representation at masked position m, $\mathbf{q}_m$ is the target quantised vector, κ is a learned logit scale, and j iterates over all candidates in the negative set.

Contrastive loss ($\mathcal{L}_{\text{cont}}$): A second InfoNCE loss is computed between the pooled fused embedding of the clean audio and the pooled acoustic embedding of the augmented view, encouraging the model to learn augmentation-invariant global representations critical for noisy channel robustness.

Codebook diversity loss ($\mathcal{L}_{\text{div}}$): To prevent codebook collapse where a small subset of entries dominates all assignments, we apply log-entropy scaling:

$$\mathcal{L}_{\text{div}} = \log K - \log \hat{\pi} \tag{4}$$

[2]Accessed on April, 2026. https://huggingface.co/facebook/w2v-bert-2.0,

[3]Accessed on April, 2026. https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

where K is the codebook size and $\hat{\pi}$ is the empirical per-batch codebook perplexity (the exponential of the assignment entropy). This formulation assigns maximum penalty when all probability mass concentrates on a single code and zero penalty when the distribution is perfectly uniform.

Auxiliary classification losses ($\mathcal{L}_{\text{aux}}$): Speaker identity, language identity, and gender classification heads are attached to the fused representation and trained with cross-entropy:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{spk}} + \mathcal{L}_{\text{lang}} + \mathcal{L}_{\text{gen}} \tag{5}$$

These losses are linearly warmed up via a weight $w_{\text{aux}}(s) = \min(1, \max(0, (s - s_0)/5000))$ where $s_0 = 25{,}000$ steps, ensuring that auxiliary objectives are introduced only after the primary acoustic representations have stabilised. This design prevents the speaker-discriminative heads from interfering with early phonetic learning.

Quantiser: A Gumbel-Softmax VQ layer with temperature annealing from $\tau_0 = 2.0$ to $\tau_\infty = 0.5$ with exponential decay at rate 5×10^{-5} per step provides the discrete target codes for the MLM loss. Annealing τ progressively sharpens the discrete assignments, improving codebook utilisation over the course of training.

Optimiser and Scheduler: AdamW ($\eta = 2 \times 10^{-4}$, weight decay $\lambda = 0.01$) with 10,000 warmup steps and cosine annealing. Gradient accumulation over 32 steps gives an effective batch size of 128. Gradient clipping at 1.0. FP16 mixed precision via `GradScaler`.

4.2 Phase 2: Convolutional Modality Alignment

After Phase 1, the speech encoder is frozen. A **Convolutional Subsampler** ($\mathcal{S}$) is trained to project fused acoustic representations into the 2048-dimensional token embedding space of the text backbone.

Subsampler Architecture: The subsampler consists of two sequential 1D convolutional layers each with stride 2, kernel size 5, and GELU activation, achieving a $4\times$ temporal compression:

$$T_{\text{out}} = \left\lfloor \frac{\lfloor \frac{T_{\text{in}}-1}{2} \rfloor - 1}{2} \right\rfloor + 1 \tag{6}$$

Output frames are capped at $T_{\text{out}} \leq 400$ tokens to bound LLM context. A final linear layer maps from the subsampler output dimension to $d_{\text{llm}} = 2048$ (Gemma-4-E2B's embedding dimension).

Prefix Stitching: The subsampled audio embeddings ($\mathbf{A} \in \mathbb{R}^{T_{\text{out}} \times 2048}$) are concatenated with the tokenised instruction prompt ($\mathbf{P}$) and target text tokens ($\mathbf{Y}$) along the sequence dimension before being passed to the frozen LLM:

$$\mathbf{X} = [\mathbf{A} \parallel \mathbf{P} \parallel \mathbf{Y}] \in \mathbb{R}^{(T_{\text{out}}+|\mathbf{P}|+|\mathbf{Y}|) \times 2048} \tag{7}$$

The LLM processes $\mathbf{X}$ causally with a standard next-token prediction head. The training loss is computed only over the $|\mathbf{Y}|$ target token positions, masking both audio and prompt positions:

$$\mathcal{L}_{\text{align}} = -\sum_{i=1}^{|\mathbf{Y}|} \log P\big(y_i \mid \mathbf{X}_{<T_{\text{out}}+|\mathbf{P}|+i}\big) \tag{8}$$

This formulation treats the projected audio sequence as a soft-prompt prefix that conditions the LLM without requiring any modification to the LLM's attention or positional encoding mechanisms.

Optimiser and Scheduler: AdamW ($\eta = 3 \times 10^{-4}$, $\lambda = 0.01$), 5% linear warmup of total steps followed by cosine annealing to $\eta_{\min} = 10^{-6}$. Effective batch size 128 (4×32 accumulation steps).

4.3 Phase 3: MLP LoRA Adapter for S2TT

Gemma-4 PLE Architecture: Gemma-4 employs PLE, a mechanism that injects learned per-layer scaling tensors by extracting raw weight matrices inside the model's forward pass. This architectural choice means that any module wrapping applied by PEFT to the attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) is bypassed. The model's internal computation path extracts `weight` tensors directly, routing around the PEFT forward override entirely. The result is that LoRA delta matrices $\Delta W = BA$ are never added to the activations, gradient flow into A and B collapses to zero, and training produces a mathematical flatline in loss with no BLEU improvement across epochs.

MLP-Targeted LoRA: We handled this by applying LoRA exclusively to the feed-forward MLP modules, whose `gate_proj`, `up_proj`, and `down_proj` projections are standard `Linear` layers not wrapped by the PLE mechanism:

$$W'_\ell = W_\ell + \frac{\alpha}{r} \cdot B_\ell A_\ell, \quad B_\ell \in \mathbb{R}^{d \times r},\ A_\ell \in \mathbb{R}^{r \times k} \tag{9}$$

with rank $r = 16$, scaling $\alpha = 32$, dropout $p = 0.05$, applied across all L layers of the Gemma-4 LLM backbone. The MLP layers are responsible for the factual knowledge and vocabulary-mapping components of LLM computation; updating them teaches the model a new audio-to-English token mapping without disturbing the PLE-sensitive attention pathway.

The trainable parameter count introduced by LoRA is:

$$N_{\text{LoRA}} = L \times 3 \times (d \cdot r + r \cdot k) \qquad (10)$$

For Gemma-4-E2B with $L = 26$ layers, $d = 2048$, $k = 8192$, $r = 16$, this yields approximately $26 \times 3 \times (32{,}768 + 131{,}072) \approx 12.9$M trainable parameters, versus $\approx$ 2B total parameters in the frozen backbone which is less than 0.65% of the full model.

All components except the LoRA matrices remain frozen during Phase 3 which includes the speech encoder, the convolutional subsampler, the linear projection, and all Gemma-4 base weights.

Instruction Prompting: Each training sample is wrapped with a structured instruction prompt encoding the source and target language:

```
Translate this audio exactly from
{source_lang} to {target_lang}.
```

The prompt and target translation are tokenised using the Gemma-4 tokeniser without additional special tokens, and the EOS token is appended to each target sequence.

Optimiser and Scheduler: AdamW ($\eta = 10^{-4}$, $\lambda = 0.01$), 5% linear warmup plus cosine annealing to $\eta_{\min} = 10^{-6}$, effective batch size 128 kept via gradient accumulation and early stopping applied with patience 5 on validation BLEU.

5 Experiments and Results

All experiments were conducted on a single NVIDIA L40s GPU with 40 GB VRAM. Each training epoch on the full 37K-sample dataset takes approximately between 1 to 1 hour 45 minutes across all three phases.

5.1 Phase 1: Speech Encoder Training

Table 2 reports selected validation loss and perplexity (PPL) checkpoints across the Phase 1 epochs. The encoder converges steadily from a validation loss of 3.537 at Epoch 0 to 0.651 at Epoch 18. PPL values in the range between 300-325 indicate healthy codebook utilisation with well-spread assignment distributions, consistent with the effectiveness of the log-entropy codebook diversity loss.

Epoch	Train Loss	Val Loss	PPL
0	4.311	3.537	1013.2
1	3.503	3.115	907.1
2	3.164	2.677	630.3
3	2.716	2.192	111.7
4	2.278	1.755	301.6
5	2.021	1.465	294.7
6	1.751	1.296	298.6
7	1.596	1.192	287.0
8	1.470	1.107	304.4
9	1.360	1.065	306.0
10	1.248	0.974	307.3
11	1.260	0.962	297.0
12	1.167	0.911	309.1
13	1.114	0.859	312.0
14	1.061	0.835	314.6
15	0.994	0.770	317.8
16	1.001	0.734	318.3
17	0.954	0.682	320.9
18	0.850	0.652	323.8

Table 2: Phase 1 speech encoder training progression.

The sharp PPL drop from 1013.2 (Epoch 0) to 111.7 (Epoch 3) reflects the warm-up of the auxiliary classification heads and the codebook settling from a near-random initialisation. After Epoch 3, PPL stabilises in the 287-324 range, indicating that the codebook has reached a healthy equilibrium without collapse.

5.2 Phase 2: Modality Alignment

Table 3 shows the validation loss descends monotonically from 3.019 to 2.877 over 11 epochs with no spikes, confirming that gradient flow through the subsampler is stable and that the LLM now receives the full audio token sequence.

Epoch	Val Loss (Fixed)
0	3.019
1	2.967
2	2.950
3	2.939
4	2.922
5	2.915
6	2.901
7	2.897
8	2.888
9	2.883
10	2.880
11	2.877

Table 3: Phase 2 alignment validation loss

5.3 Phase 3: S2TT LoRA Fine-Tuning

Table 4 presents the full Phase 3 training trajectory. The MLP LoRA adapter drives extremely rapid convergence: validation loss drops from 2.192 to 0.337 over 8 epochs and proxy BLEU rises from 19.26 to 91.21 in the same span. Both metrics plateau at Epoch 10 (BLEU 91.29, Val Loss 0.344), after which the system begins to slightly overfit (BLEU 91.27 at Epoch 11), triggering the patience counter.

Epoch	Val Loss	Proxy BLEU
0	2.192	19.26
1	1.122	52.80
2	0.603	77.33
3	0.443	**85.83**$\star$
4	0.385	88.98
5	0.365	90.01
6	0.350	90.73
7	0.338	91.13
8	0.337	91.21
9	0.340	91.27
10	0.344	**91.29**
11	0.350	91.27

Table 4: Phase 3 S2TT LoRA validation results. Submitted checkpoints marked in bold with $\star$ indicates primary run and without $\star$ indicates the contrastive run.

The steep BLEU gain in Epochs 0-3 (19→86) reflects the MLP layers rapidly learning the acoustic to text mapping. The shallower gain in Epochs 4-10 (88→91) reflects fine-grained vocabulary alignment and correction of tonal language boundary errors. The slight val-loss uptick at Epoch 9 while BLEU continues to improve (91.27→91.29) indicates that the model is shifting probability mass to longer or lower-frequency tokens that contribute positively to n-gram overlap but marginally increase the cross-entropy.

All reported BLEU scores are *proxy teacher-forced* BLEU, computed by slicing the highest-probability token at each target position under teacher-forced decoding rather than autoregressive generation. This is a valid convergence diagnostic but is systematically higher than free-decoding BLEU. Submitted system outputs were generated via autoregressive beam-search decoding.

5.4 Final Evaluation Results

Table 5 presents the official autoregressive beam-search decoding results for AURA-ST across the three African source languages, evaluated using SpBLEU, ChrF, and SSA-COMET.

Language	SSA-COMET	SpBLEU	ChrF
Yo–En	0.57	19.5	41.0
Ha–En	0.34	5.2	23.2
Ig–En	0.21	4.6	16.7

Table 5: Final test-set performance of AURA-ST measured using SpBLEU, ChrF, and SSA-COMET. Yo–En, Ha–En, and Ig–En denote Yoruba→English, Hausa→English, and Igbo→English, respectively.

The empirical results reveal a stark contrast between free-decoding performance and the proxy teacher-forced validation metrics reported during development, which reached a proxy BLEU of 91.29. This discrepancy highlights a significant exposure-bias bottleneck inherent to the proposed prefix-stitching architecture. While the MLP-targeted LoRA adapter successfully aligns the audio representations with the text space under teacher-forced supervision, the frozen Gemma-4-E2B backbone struggles to maintain stable generation when conditioned solely on its own historical outputs during autoregressive inference.

The system exhibits a highly uneven performance profile across the three language pairs. Yoruba→English emerges as the strongest configuration, achieving an SpBLEU of 19.5, a ChrF score of 41.0, and an SSA-COMET score of 0.57. These results support the motivation behind the proposed dual-stream encoder design. The auxiliary ResNet34 branch explicitly models F0 contours and durational characteristics that are absent from purely phonemic representations. For Yoruba, which employs a three-level lexical tone system, these acoustic cues appear to preserve semantic information sufficiently for successful downstream decoding and contribute to the substantially stronger translation quality observed on the official test set.

In contrast, Hausa→English (5.2 SpBLEU, 23.2 ChrF, 0.34 SSA-COMET) and Igbo→English (4.6 SpBLEU, 16.7 ChrF, 0.21 SSA-COMET) suffer substantial degradation in both lexical overlap and semantic adequacy. Although Hausa and Igbo are also tonal languages, the current architecture relies solely on a natural-language instruction prompt without explicit token-level language conditioning. Consequently, the shared multimodal embedding space likely experiences cross-lingual interference, causing the acoustic prefixes for these languages to anchor less effectively within the LLM token manifold. During autoregressive decoding, this weak

alignment can lead to rapid divergence, hallucination, and loss of translation fidelity, resulting in significantly lower SpBLEU and semantic evaluation scores compared to Yoruba.

5.5 Ablation Observations

The development process yielded two natural ablations:

Without warmup (Phase 2, initial run): training loss oscillated ± 0.3 between epochs due to high initial LR hitting the randomly initialised subsampler. Adding 5% linear warmup eliminated all oscillations and produced smooth descent.

With attention LoRA (Phase 3, rejected run): applying LoRA to `q_proj`/`k_proj`/`v_proj` under Gemma-4 PLE produced zero BLEU improvement across 3 epochs (BLEU ≈ 0.0, Val Loss static at 2.2). Switching to MLP injection produced BLEU 19.26 at Epoch 0 and 52.80 at Epoch 1.

6 Discussions

Dual-Stream Complementarity: The paralinguistic ResNet34 stream captures F0 contours and speaker-specific durational patterns that the purely phonemic w2v-BERT representation does not model at the frame level. This is particularly important for Yoruba, which has a three-level lexical tone system where minimal pairs are distinguished solely by pitch, and for Igbo, which uses a two-level tone system.

Attention-Free Prefix Architecture: By treating audio as a prefix rather than via cross-attention, AURA-ST avoids the quadratic complexity of cross-attention over long audio sequences. With the $4\times$ subsampling and 400-token cap, a 30-second utterance at 50 frames/second produces at most $\min(375, 400) = 375$ audio tokens. At total context length 800 (400 audio + 200 prompt + 200 target), the causal self-attention in Gemma-4 operates within its standard context window without any architectural modification, making the system straightforward to scale to larger LLMs.

PLE-Aware Adapter Injection: The identification of the Gemma-4 PLE incompatibility with attention LoRA is a practically significant finding for any PEFT-based multimodal adaptation of Gemma-4. The root cause is that PLE bypasses standard module dispatch by accessing `weight` attributes directly in the forward computation, which is the entry point PEFT wraps. MLP layers in Gemma-4 do not use PLE and therefore remain PEFT-compatible. We recommend that future work applying LoRA or other wrapper-based parameter-efficient methods to Gemma-4 variants verify gradient flow through the wrapped layers explicitly before committing to a target module configuration.

Limitations: The proxy BLEU metric used during training overstates real-world translation quality but it is expected that autoregressive beam-search BLEU scores on the official test set will be lower. The current system does not include any language-ID conditioning beyond the free-text instruction prompt, which may cause confusion for closely related dialect clusters within the three source languages. Additionally, the Phase 1 encoder was trained only on the IWSLT 2026 Track 1 corpus where a larger multilingual pretraining corpus could improve generalisation to unseen acoustic conditions.

7 Conclusion

We presented AURA-ST, a three-phase modular speech translation pipeline that connects a dual-stream acoustic encoder to a frozen Gemma-4-E2B language model via a convolutional subsampler and MLP-targeted LoRA adapters, submitted to IWSLT 2026 Track 1 for African low-resource speech translation. The system demonstrates strong convergence with Phase 1 encoder validation loss 0.651, Phase 2 alignment val loss 2.877, and Phase 3 proxy BLEU 91.29 on the validation set.

Three engineering contributions emerged from the development process with broader applicability. It includes a learning-rate warmup protocol for training randomly initialised projection modules adjacent to frozen encoders and the identification of Gemma-4's PLE architecture as incompatible with standard attention-level PEFT injection and its resolution via MLP-targeted LoRA.

Future work will focus on autoregressive decoding tuning, cross-lingual adapter sharing across the three African source languages, extension to the Celtic track, and ablating the contribution of each dual-stream component on per-language test-set BLEU.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Se-

bastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, and 45 others. 2023. SeamlessM4T: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–251. IEEE.

Yassir Fathullah, Chunxi Liu, Egor Lakomkin, Ke Li, Yuan Shangguan, Ozlem Kalinli, Christian Fuegen, Jagadeesh Balam, Boris Ginsburg, and Mike Seltzer. 2024. Prompting large language models with speech recognition abilities. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8370–8374. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 827–838. Association for Computational Linguistics.

Marie Maltais, Yejin Jeon, Min Ma, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Maryam Ibrahim Mukhtar, Daud Abolade, Joel Okepefi, Johnson Sewedo, and David Ifeoluwa Adelani. 2026. Naijas2st: A multi-accent benchmark for speech-to-speech translation in low-resource nigerian languages. *arXiv preprint arXiv:2604.16287*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518. PMLR.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7409–7421. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pages 776–794. Springer, Cham.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of Interspeech 2017*, pages 2177–2181. ISCA.

Haibin Xie, Zach Evans, Chunxi Liu, Ruizhe Huang, Modan Tailleur, Yassir Fathullah, Yunfei Chu, Xinfa Zhu, Wei Li, and Lei Xie. 2024. LLaSA: Large language and speech assistant. *Preprint*, arXiv:2501.03679.

Hossein Zeinali, Lukáš Burget, Johan Rohdin, Oldřich Plchot, Ondrej Glembek, Oldrich Plchot, and Jan Černocký. 2019. BUT system description to VoxCeleb speaker recognition challenge 2019. *Preprint*, arXiv:1910.12592.