Benchmarking Multimodal Semantic Alignment Between Speech and Text Representations in the Igbo Language

Chidiebere Christopher v**chidiebere.vc@gmail.com**
Independent researcher

Abstract

We present the first systematic benchmark for cross-modal semantic alignment between speech and text in Igbo—a tonal Niger-Congo language spoken by ≈45 million people with virtually no prior multimodal NLP representation. Using the WAXAL corpus, we extract 699 stratified utterance pairs and encode them with Whisper-tiny (speech, 39M parameters, 384-d) and `paraphrase-multilingual-MiniLM-L12-v2` 118M parameters, 384-d). Zero-shot cross-modal cosine similarity is −0.0009, statistically indistinguishable from random, confirming a complete alignment gap. A lightweight linear projection (147,840 parameters) trained with symmetric InfoNCE and similarity-based hard negative mining with curriculum re-mining achieves Speech-to-Text Recall@1 of 0.0658 (5.1×), Recall@10 of 0.3362 (3.0×), and MRR of 0.1557 (2.7×) over the zero-shot baseline in a 100-candidate pool. Text-to-Speech retrieval achieves R@10 of 0.3777 (3.0×). Alignment is statistically significant ($t = 15.95$, $p = 2.81 \times 10^{-54}$). Error analysis identifies utterance duration ($r = -0.19$) and word count ($r = +0.29$) as systematic failure predictors. We release all embeddings, evaluation code, and benchmark protocols.

Keywords: Igbo, multimodal alignment, speech-text retrieval, contrastive learning, low-resource NLP, African languages, WAXAL

1. Introduction

Multimodal speech-text alignment has advanced dramatically in high-resource languages through models such as Whisper (Radford et al., 2022), SeamlessM4T (Barrault et al., 2023), and CLIP (Radford et al., 2021). Yet this progress is concentrated in a narrow set of languages. For African languages—2,000+ languages spoken by 1.4 billion people—cross-modal alignment research is effectively absent (**?**Osei et al., 2023). Igbo is a particularly urgent case: 45 million speakers, a tonal phonology in which suprasegmental tone patterns carry lexical meaning, a diacritic-augmented Latin orthography (*ị, ọ, ụ*) frequently stripped in digital text, and zero published cross-modal alignment benchmarks.

The core question this paper addresses is empirical: do off-the-shelf multilingual speech and text encoders produce geometrically compatible representations for Igbo, and if not, can lightweight alignment training bridge the gap under the data scarcity and compute constraints characteristic of African NLP research? We make four contributions:

C1: First cross-modal alignment benchmark for Igbo using WAXAL (Diack et al., 2026), with all embeddings and evaluation code released.

C2: First empirical zero-shot baseline—paired cosine similarity of −0.0009, confirming a complete alignment gap between Whisper-tiny and multilingual MiniLM for Igbo.

C3: Lightweight linear projection with InfoNCE + hard negative mining achieving 5.1× R@1 improvement, trained in 1.6 seconds on CPU.

C4: Systematic failure mode analysis identifying utterance duration, semantic specificity, and proper noun density as primary predictors of alignment failure.

2. Related Work

2.1. Multilingual Speech and Text Encoders

Whisper (Radford et al., 2022), trained on 680,000 hours across 99 languages, produces encoder representations shown to carry abstract semantic properties in upper layers (Shon et al., 2023). MMS (Pratap et al., 2023) extends coverage to 1,100+ languages including Igbo ASR. For text, `paraphrase-multilingual-MiniLM-L12-v2` (Wang et al., 2022a) achieves efficient multilingual sentence encoding via knowledge distillation across 50+ languages. Boito et al. (2022) show cross-modal alignment quality degrades systematically with typological distance from pretraining-dominant languages—directly predicting the near-zero zero-shot alignment we observe for Igbo. Adelani et al. (2022) and Alabi et al. (2022) document that even state-of-the-art multilingual models require targeted adaptation to achieve acceptable performance on Igbo NLP tasks.

2.2. Cross-Modal Alignment and Contrastive Learning

InfoNCE contrastive loss (Chen et al., 2021) has become the dominant objective for speech-text alignment, maximising mutual information between paired representations. Shih et al. (2023) demonstrate cross-modal speech-text retrieval via SpeechCLIP; Zhao et al. (2022) show symmetric InfoNCE outperforms asymmetric variants for spoken language understanding. Robinson et al. (2021) establish theoretically and empirically that hard negative mining—using the most similar wrong candidates rather than random ones—yields substantially stronger contrastive learning signal. Tian et al. (2022) demonstrate curriculum re-mining adapts negative difficulty to the current model state, maintaining training signal throughout. No prior work applies these methods to any West African tonal language.

2.3. African Language Benchmarking

AfriSpeech-200 (Olatunji et al., 2023) benchmarks African-accented English ASR, revealing systematic commercial system degradation. FLEURS (Conneau et al., 2022) covers 102 languages including Igbo (`ig_ng`) for ASR evaluation. MasakhaNER 2.0 (Adelani et al., 2022) extends NLP benchmarking to 20 African languages. WAXAL (Diack et al., 2026)—the data source for this paper—provides the first studio-quality Igbo TTS recordings at scale (1,911 utterances, Media Trust Nigeria, CC-BY-4.0), making cross-modal alignment research feasible for the first time.

3. Methodology

3.1. Dataset and Preprocessing

We load all four WAXAL Igbo TTS parquet files (`ibo-train` ×2, `ibo-validation`, `ibo-test`) comprising 1,911 raw records across 8 speakers with near-perfect gender balance (50.3% F / 49.7% M). Audio is stored as embedded MP3 bytes (48 kHz native) decoded in-memory via `librosa` without disk writes. Validation filters exclude records shorter than 0.5s or longer than 30s, yielding 1,343 valid records (568 excluded, 29.7%). We draw a stratified subsample of 699 records balanced across three duration buckets (33rd/66th percentile splits) using `SEED = 42`. Audio is resampled to 16 kHz, peak-normalised to $[-1, 1]$, and truncated at 30s. Text undergoes Unicode NFC normalisation (critical for Igbo diacritics), artifact removal, whitespace collapse, and lowercasing—diacritics are preserved throughout.

Figure 1. WAXAL Igbo TTS subsampled corpus (n=699): audio duration (mean=12.2s), transcription length (mean=14.0 words), speaker gender balance.

3.2. Encoders and Embedding Extraction

Speech embeddings are extracted from the frozen Whisper-tiny encoder (`openai/whisper-tiny`, 39M params, 4 transformer layers, d=384). Input: log-mel spectrogram padded to exactly 3,000 frames (30s at 16 kHz) by `WhisperFeatureExtractor`. Output: mean-pool of 1,500 last-layer hidden states → 384-d vector → L2-normalise. Text embeddings use `paraphrase-multilingual-MiniLM-L12-v2` (118M params, 12 layers, d=384) via Sentence-Transformers with `normalize_embeddings=True`. Both encoders produce identically-sized 384-d unit-norm vectors, making a square linear projection $W_p \in \mathbb{R}^{384\times384}$ sufficient for alignment—147,840 total trainable parameters. All embeddings are extracted once and cached as `.npy` files; subsequent experiments operate on cached matrices only.

3.3. Linear Projection and Contrastive Training

The projection $f_\theta(s) = \text{Norm}(\text{Dropout}(\text{LayerNorm}(W_p \cdot s + b_p)))$ maps speech embeddings into the text embedding space. Training uses symmetric InfoNCE loss at temperature $\tau = 0.07$, simultaneously optimising speech→text and text→speech alignment. Hard negative mining computes the full 699×699 zero-shot similarity matrix and identifies the top-5 most similar wrong text candidates per sample. Two hard negatives are injected per batch element, expanding the 32-sample batch to 32 speech queries against 96 text candidates with explicit ground-truth labels. Hard negatives are re-mined from the projector's current similarity matrix every 7 epochs (curriculum re-mining). Optimiser: AdamW (lr $= 3 \times 10^{-4}$, weight decay $= 10^{-4}$) with cosine annealing; gradient clipping at 1.0. Train/val split: 80/20 (559/140 pairs). Total training time: 1.6 seconds on CPU.

3.4. Evaluation Protocol

Two retrieval tasks: Speech→Text (S→T) and Text→Speech (T→S). Each query is ranked against a pool of 100 candidates (1 correct + 99 random distractors, fixed seed). Metrics: Recall@K ($K = 1, 5, 10$), Mean Reciprocal Rank (MRR), Mean/Median Rank, and Paired

Cosine Similarity. All metrics reported with 95% bootstrap confidence intervals (1,000 resamples). Statistical significance of alignment assessed via one-sided Welch's t-test comparing 500 paired vs. 1,500 unpaired cosine similarities. Random baseline: R@1= 0.010, R@5= 0.050, R@10= 0.100, MRR≈ 0.051.

4. Results

4.1. Zero-Shot Baseline

The zero-shot paired cosine similarity of -0.0009 is statistically indistinguishable from the expected value of zero for random unit vectors in $\mathbb{R}^{384}$. Mean rank of 49.47 (S→T) and 46.44 (T→S) in a pool of 100 confirm that the correct answer is buried at the candidate list midpoint. Zero-shot R@1 of 0.0129 is barely above random (0.010). Figure 2 visualises this: the similarity heatmap shows no visible diagonal structure—ground-truth pairs are indistinguishable from random pairs.

Figure 2. Zero-shot cosine similarity matrix (first 60 pairs). No visible diagonal—paired speech-text embeddings are geometrically indistinguishable from random.

4.2. Training Dynamics

Train loss drops from 3.86 to 3.30 in epochs 1–5. Re-mining events at epochs 8 and 15 produce characteristic spikes (to 3.68 and 4.80 respectively) as the model encounters harder negatives, then recover—the curriculum contrastive learning signature (Tian et al., 2022). Best validation loss: 4.5985. Figure 3 shows the full trajectory.

Figure 3. InfoNCE loss trajectory across 20 epochs. Loss spikes at epochs ≈7 and ≈14 correspond to curriculum re-mining events introducing harder negatives.

4.3. Retrieval Benchmark

Table 1 reports full benchmark results.

Table 1: Full benchmark results (pool size=100, $N = 699$). Random baseline: R@1=0.010, R@5=0.050, R@10=0.100, MRR≈0.051.

Method	Direction	R@1	R@5	R@10	MRR	MeanRank	Improvement
Zero-shot	S→T	0.0129	0.0615	0.1116	0.0580	49.47	—
Zero-shot	T→S	0.0129	0.0629	0.1259	0.0616	46.44	—
Projected	S→T	0.0658	0.2146	0.3362	0.1557	25.83	×5.1 R@1
Projected	T→S	0.0644	0.2332	0.3777	0.1638	23.31	×5.0 R@1

The projection yields 2.7×–5.1× improvements across all metrics and both retrieval directions. R@10 of 0.3362 (S→T) and 0.3777 (T→S) indicate that the correct match appears in the top 10 for roughly 1-in-3 queries. Median rank improves from 49→19 (S→T) and 45→16 (T→S). T→S consistently outperforms S→T at higher recall thresholds,

attributed to greater intra-class consistency of MiniLM text embeddings. Confidence intervals for zero-shot and projected R@1 do not overlap, confirming statistical reliability.

Figure 4. Complete benchmark results. Zero-shot performance is near-random; projection yields consistent 2.7×–5.1× improvements across all metrics and both retrieval directions.

4.4. Statistical Validation

One-sided t-test comparing 500 paired vs. 1,500 unpaired projected cosine similarities: $t = 15.95$, $p = 2.81 \times 10^{-54}$. Paired mean = 0.112, unpaired mean ≈ 0.049. The overwhelming significance rules out degenerate transformation—the projection has learned genuine cross-modal semantic structure.

Figure 5. Paired (blue, n=500) vs. unpaired (red, n=1,500) cosine similarity distributions after projection. Paired mean= 0.112 > unpaired mean= 0.049 ($t = 15.95$, $p = 2.81 \times 10^{-54}$).

5. Discussion

5.1. Why the Zero-Shot Gap is Complete

Three compounding mechanisms explain the −0.0009 zero-shot similarity. First, *pretraining objective mismatch*: Whisper's representations optimise acoustic discriminability; MiniLM's optimise semantic equivalence—these objectives are mathematically orthogonal. Second, *Igbo underrepresentation*: with negligible Igbo content in either model's training data, whatever cross-modal structure exists for high-resource languages is absent for Igbo. Third, *tonal and diacritic information loss*: Igbo's lexical tones are encoded acoustically but largely absent from digital text; its diacritics are systematically stripped in online corpora, creating bidirectional signal loss unique to tonal diacritic-orthography languages. This confirms the alignment failure is qualitative—rooted in Igbo's specific phonological properties—not merely a quantitative data shortage.

5.2. Failure Mode Analysis

Duration–rank correlation ($r = -0.19$): short clips (3–5s) underperform because silence-padding dilutes mean-pooled Whisper embeddings. Word count–similarity correlation ($r = +0.29$): longer sentences produce more semantically specific MiniLM embeddings, geometrically more distinct from distractors. Worst-case retrieval (ranks 92–99) clusters around proper noun-dense sentences (Goodluck Jonathan, Yar'adua, 3rd Mainland Bridge)—both encoders fail for Nigerian entities absent from their training data, but fail in uncorrelated ways that projection cannot align. A hub phenomenon (single generic sentence repeatedly retrieved as a false match) confirms the need for CSLS post-processing in future work.

Figure 6. Failure analysis: duration vs. rank ($r = -0.192$), word count vs. paired similarity ($r = 0.288$), rank by gender (no systematic effect).

5.3. Broader Implications

Three findings carry implications beyond Igbo. First, zero-shot alignment cannot be assumed for African languages: even state-of-the-art multilingual models fail completely for

Igbo. Second, lightweight alignment is viable under African research constraints: 147,840 parameters, 559 training pairs, 1.6s on CPU produce statistically significant and practically meaningful improvements—no GPU required. Third, tonal language properties require explicit architectural attention: scaling data alone will not close the gap without tone-aware acoustic features, diacritic-preserving tokenisation, and cross-modal training objectives that model tone-meaning correspondence.

6. Limitations and Future Work

The key limitations and their direct remedies are:

Data scope: 8 speakers, studio TTS only, 568 records lost to duration filters. Remedy: integrate FLEURS `ig_ng` (30+ speakers, natural speech) and chunk long audio instead of discarding it, recovering ≈300 additional samples.

Encoder choice: Whisper-tiny (4 layers) is the weakest Whisper variant; MiniLM is a distilled compact model. Remedy: systematic comparison with Whisper-base/small and LaBSE/mE5; evaluate Meta MMS encoder (explicit Igbo ASR supervision) as the most promising alternative speech encoder.

Architecture: Linear projection cannot model heterogeneous alignment difficulty. Remedy: two-layer MLP, cross-attention projection, or Houlsby adapter fine-tuning of encoder layers.

Evaluation: Pool size 100 underestimates full-corpus difficulty. Remedy: report full-corpus metrics ($N = 699$) alongside pool-based, apply CSLS hubness reduction, develop a hard-negative evaluation subset.

Linguistic depth: No analysis of tonal complexity, morphological density, or diacritic effects. Remedy: tone-annotated evaluation subsets, tonal probing classifiers on Whisper hidden states, collaboration with Igbo linguistic experts.

7. Conclusion

We have established the first cross-modal alignment benchmark for Igbo, demonstrating that off-the-shelf multilingual encoders produce geometrically orthogonal speech and text representations for the language (zero-shot cosine similarity = -0.0009), and that a minimal-parameter projection trained with InfoNCE contrastive loss and hard negative mining can bridge a substantial portion of this gap—achieving R@10 of 0.3362 (S→T) and 0.3777 (T→S) with overwhelming statistical significance ($p = 2.81 \times 10^{-54}$)—entirely on consumer CPU hardware. Our failure mode analysis reveals that alignment quality is systematically predicted by utterance duration and semantic specificity, with proper noun-dense Nigerian content representing the hardest cases for current architectures.

The zero-shot result is not a failure of our approach—it is the primary scientific finding. It proves that 45 million Igbo speakers are currently excluded from the benefits of speech-text AI technology not because the problem is intractable, but because the foundational measurement infrastructure did not exist. This paper provides that infrastructure: a reproducible baseline, a released benchmark, and a map of exactly where the work needs to go. The gap between -0.0009 and the perfect alignment of 1.0 is the research programme for Igbo multimodal NLP. We have measured its starting point.

References

Adebara, I., & Abdul-Mageed, M. (2022). Towards Afrocentric NLP for African languages: Where we are and where we need to go. *Proceedings of ACL 2022*, 1, 2236–2252. https://doi.org/10.18653/v1/2022.acl-long.162

Adelani, D. I., Abbott, J., Neubig, G., et al. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the ACL*, 9, 1116–1131. https://doi.org/10.1162/tacl_a_00416

Adelani, D. I., Neubig, G., Ruder, S., et al. (2022). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. *Proceedings of EMNLP 2022*, 4488–4508. https://doi.org/10.18653/v1/2022.emnlp-main.298

Alabi, J. O., Adelani, D. I., Mesham, M., & Bash, A. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. *Proceedings of COLING 2022*, 4336–4349.

Barrault, L., Chung, Y., Meglioli, M. C., et al. (2023). SeamlessM4T: Massively multilingual and multimodal machine translation. arXiv:2308.11596.

Boito, M. Z., Bentivogli, L., Gaido, M., Negri, M., & Turchi, M. (2022). Speech translation for low-resource languages: Evaluating cross-lingual transfer learning approaches. *Proceedings of INTERSPEECH 2022*, 2518–2522.

Chen, T., Kornblith, S., Sohl-Dickstein, J., Hinton, G., & Norozi, M. (2021). A simple framework for contrastive learning of visual representations. *Proceedings of ICML 2021, PMLR 139*, 1597–1607.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of INTERSPEECH 2021*, 2426–2430.

Conneau, A., Ma, M., Khanuja, S., et al. (2022). FLEURS: Few-shot learning evaluation of universal representations of speech. *Proceedings of IEEE SLT 2022*, 798–805.

Diack, A., Asare, A., Adusei, B., et al. (2026). WAXAL: A large-scale multilingual African language speech corpus. arXiv:2602.02734.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. *Proceedings of ACL 2022*, 1, 878–891.

Guzhov, A., Raue, F., Hees, J., & Dengel, A. (2022). AudioCLIP: Extending CLIP to image, text and audio. *Proceedings of ICASSP 2022*, 976–980.

Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *Proceedings of ICML 2021, PMLR 139*, 4904–4916.

Kim, J., Winata, G. I., & Fung, P. (2023). STAM: A speech-text alignment model for spoken language understanding. *Proceedings of ACL 2023*, 1, 3598–3612.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of ICLR 2019*.

Olatunji, T., Emezue, C., Nakatumba-Nabende, J., & Mayhew, S. (2023). AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the ACL*, 11, 57–73.

Osei, S., Oppong, F., Arthur, S., et al. (2023). AfroNLU: Benchmarking African language understanding. *Proceedings of EACL 2023*, 311–325.

Peng, P., & Harwath, D. (2022). Fast-slow transformer for visually grounding speech. *Proceedings of ICASSP 2022*, 8527–8531.

Pratap, V., Tjandra, A., Shi, B., et al. (2023). Scaling speech technology to 1,000+ languages. *Proceedings of ACL 2023*, 1, 4693–4710.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of ICML 2021, PMLR 139*, 8748–8763.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *Proceedings of ICML 2023, PMLR 202*, 28492–28518.

Robinson, J., Chuang, C., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. *Proceedings of ICLR 2021*.

Shih, Y., Chen, B., & Harwath, D. (2023). SpeechCLIP: Integrating speech with CLIP for cross-modal representations. *Proceedings of IEEE SLT 2023*, 1–8.

Shon, S., Pasad, A., Wu, F., et al. (2023). SLUE-PERB: A spoken language understanding performance benchmark and toolkit. *Proceedings of ICASSP 2023*, 1–5.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., & Isola, P. (2022). What makes for good views for contrastive learning. *Advances in NeurIPS 33*, 6827–6839.

Wang, L., Yang, N., Huang, X., et al. (2022). Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2022). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in NeurIPS 33*, 5776–5788.

Wang, L., Yang, N., Huang, X., et al. (2024). Multilingual E5 text embeddings: A technical report. arXiv:2402.05672.

Yi, C., Tao, J., Tian, Z., & Liu, B. (2022). Cross-lingual contrastive learning for speech representations. *Proceedings of INTERSPEECH 2022*, 3029–3033.

Zhang, S., Qian, K., Gao, H., Hou, Y., & Xu, B. (2022). Hard negative sampling strategies for contrastive representation learning in speech processing. *Proceedings of INTERSPEECH 2022*, 1826–1830.

Zhao, Y., Li, J., & Xu, B. (2022). Contrastive learning for speech-text alignment in spoken language understanding. *Proceedings of ICASSP 2022*, 7177–7181.