NOUN Interdisciplinary Journal of Computing, E-learning & Application (NOUN-IJCEA)
https://ijcea.nou.edu.ng
Volume 1 Number 1, April 2026 pp. 169-181

Development of a Bi-Lingual Speech Translator for Igbo to English Using Hubert and Differential Evolution

***Frederick Chukwuma Alu, Jumoke Falilat Ajao, Sulaiman Olaniyi Abdulsallam, Maryam Adeola Abdullahi**

Department of Computer Science, Kwara State University, Malete, Nigeria
*Corresponding Author: fredsatchuks@gmail.com; Jumoke.ajao@kwasu.edu.ng

ABSTRACT

The language barrier between Igbo and English speakers in Nigeria limits effective communication in education, governance, and social interactions. Most existing speech translation tools rely on text-based or cascaded architectures that introduce latency, and fail to capture the nuances of spoken Igbo. This paper presents a direct speech-to-speech translation (S2ST) system that converts Igbo speech into English speech without intermediate text generation. The system adopts a speech-to-unit (S2UT) framework: source Igbo audio is transformed into 80-dimensional log-mel features, while target English audio is represented as discrete acoustic units obtained by quantizing HuBERT features. A Transformer-based encoder–decoder, implemented in fairseq, learns to map the source features to target unit sequences. A HiFi-GAN vocoder then synthesizes the final English waveform. Three optimization strategies were compared: no optimizer, the Adam Optimizer, and a Differential Evolution (DE) fine-tuning applied to a subset of the decoder's output projection weights. Experiments show that the Adam-trained model achieves a BLEU4 score of 23.97, a Unit Error Rate (UER) of 66.18%, and a token-level accuracy of 33.69%. The model without any optimizer fails to converge (BLEU4 = 0.15). though the DE fine-tuning reduced the loss by a little margin but still yields identical performance, confirming that gradient-based training already reaches a strong local optimum.

Keywords: Speech-to-speech translation, Igbo language, HuBERT Model, Discrete units, Differential Evolution, Low-resource language

1. INTRODUCTION

Nigeria is home to over 500 languages, making cross-linguistic communication a major challenge. The Igbo language, spoken by more than 27 million people, is one of the country's major languages, but it remains under-resourced in natural language processing (NLP). The lack of effective translation tools between Igbo and English hinders education, fosters ethnic misunderstandings, and limits access to information (Nwonu et al., 2023; Omachonu, 2012).

Early machine translation systems for African languages were rule-based or statistical (Abiola et al., 2015; Lopez, 2008), but they struggled with the rich morphology and tonal nature of Igbo. Some neural approaches, particularly transformer-based models (Vaswani et al., 2017) and self-supervised speech representations like HuBERT (Hsu et al., 2021), have opened new possibilities for low-resource speech translation. Direct speech-to-speech translation (S2ST) using discrete acoustic units (Lee et al., 2022a; Li et al., 2022) has shown promise by avoiding text as an intermediate representation, thereby preserving prosody and reducing error propagation.

©2026 by the author(s)

This article is open access under the CC BY-NC license
https://creativecommons.org/licenses/by-nc-nd/4.0/

However, most existing S2ST systems are designed for high-resource languages and rely heavily on gradient-based optimization. Evolutionary algorithms such as Differential Evolution (DE) (Storn & Price, 1997) have been used in other domains to escape local optima, but their application to S2ST remains unexplored. This paper addresses these gaps by:

i. Designing a direct S2ST system from Igbo to English using a speech-to-unit (S2UT) architecture with HuBERT discrete units.
ii. Comparing three optimization strategies: no optimizer, Adam (gradient-based), and DE (Evolution – based) fine-tuning of a subset of the decoder's output projection weights.
iii. Evaluating the system on a self-collected parallel Igbo–English speech corpus using Unit Bi-Lingual Evaluation Understudy (BLEU4), Unit Error Rate (UER), and Token-level accuracy.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the methodology, including data collection, preprocessing, model architecture, and optimization. Section 4 presents the experimental results and discussion. Section 5 concludes and outlines future work.

2. LITERATURE REVIEW

The evolution of S2ST for African languages has progressed from cascaded text-based systems to end-to-end speech-to-unit models. Abbott & Martinus (2018) demonstrated that a Transformer model significantly improves English-to-Setswana translation, showing the viability of neural architectures for African languages. Jia et al. (2019) attempted direct S2ST but reported poor voice transfer; their later work (Jia et al., 2022) achieved voice preservation using a Conformer encoder and neural vocoder, though real-time performance remained limited.

A major breakthrough came with textless S2ST using discrete speech units (Lee et al., 2022a; Li et al., 2022). Lee et al. (2022a) showed that discrete units are easier to model than continuous mel-spectrograms. Popuri et al. (2022) demonstrated that self-supervised pre-training (e.g., wav2vec 2.0) yields BLEU gains of 6.6–12.1. For low-resource language pairs, Rashidi & Sameti (2025) built a Persian–English direct S2ST system using HuBERT units and synthetic data, achieving a BLEU score of 18.4. Most relevant to this research, Maltais et al. (2026) benchmarked cascaded, end-to-end, and AudioLLM approaches for Nigerian languages (Hausa, Yorùbá). They found that cascaded pipelines suffer from error propagation and that lexical metrics like ChrF are unreliable for diacritic-rich languages. Their results underscore the need for custom, direct S2ST models for Igbo.

Limited or no previous study has applied Differential Evolution to fine-tune a HuBERT-based S2ST model, nor has a direct Igbo-to-English S2ST system been reported. This work fills that gap while Table 1 presents a review of some related works carried out in the area of Speech-to-Speech translation task.

Table 1: Review of past related work

Author(s) and Year	Paper title	Methodology	Algorithm	Results	Remark
Abbott, & Martinus, (2018)	Towards Neural Machine Translation for African	Convolutional sequence-to-sequence combined with Transformer, to compare to existing phrase-based SMT for English-to-Setswana translation research.	Vector Quantized Variational Autoencoder (VQ-VAE) ConvS2S, Tensor2tensor Transformer, phrase-based English-Setswana system	ConvS2S BLEU score of 27.77, with 1.02 BLEU points below phrase-based and Transformer model 33.12 with 5.33 BLEU points compared to Phrase-based English-to-Setswana translation.	The study confirms that the use of NMT systems, in particular the Transformer model, can improve the state of the art in English to Setswana translation
Jia et al, 2019	Direct speech-to-speech translation with a sequence-to-sequence model	The neural network is trained end-to-end, allowing it to learn the mapping of speech spectrograms from one language to target spectrograms in another language, in a different voice style.	WaveRNN neural Vocoder, LSTM and dvector V3 model	The model achieves high translation quality on datasets, although performance falls below baseline cascaded method. The voice transfer did not work, reflecting the difficulty of the cross-language voice transfer task, as well as evaluation	the source and target speakers for each pair are different. Explore alternate training strategies
Jia et al, 2022	Translatotron 2: High-quality direct speech-to-speech translation with voice preservation	Made use of a transformer and CNN as encoder, layers of LSTM as decoder and a multihead Attention like transformer. A synthesizer for the speech	CNN and LSTM	It outperforms Translatotron on both translation quality (up to +15.5 BLEU) and speech generation quality and was able to preserve each speaker's voice	Does not support simultaneous translation, cross-lingual prosody transfer, and the training was not self-supervised.
Li et al., 2022	Textless Direct Speech-To-Speech Translation	used a pre-trained multilingual w2v-BERT	pre-trained multilingual w2v-BERT model,	The model can effectively handle multilingual	Identifies failure patterns in the models pronunciation

Author(s) and Year	Paper title	Methodology	Algorithm	Results	Remark
	With Discrete Speech Representation	speech encoder on dataset, a speech quantizer (VQ-VAE) to extract discrete representations from the target speech to train the decoder.	Transformer stack as the VQ-VAE decoder and the S2ST acoustic synthesizer.	speech translation Translatotron model outperformed the prior state-of-the-art textless model by a significant margin of +18.5 BLEU	of parts of the source speech like names and proper nouns. Might not be suitable for low-resource languages.
Lee et al., 2022b	Textless Speech-to-Speech Translation on Real Data	Using the Multilingual HuBERT on unlabelled speech and for Speech normalization. Apply mHuBERT to get orig. unit used to train the speech normalizer.	Multilingual HuBERT (mHuBERT), Connectionist Temporal Classification (CTC) and Speech to Unit Translation (S2UT) Model	Using reference speaker speech normalization Inproves the translation quality and reduces variations in the multi-speaker target speech while retaining the lexical content	Discreate unit output (norm-unit) from speech normalizer is a shorter and cleaner target for training S2UT models.
Popuri, et al., (2022)	Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation	Used a self-supervised pre-training with unlabelled speech data and data augmentation techniques to enhance direct speech-to-speech translation models.	Wav2vec as Encoder and Unit mBART as Decoder	Self-supervised pre-training shows a BLEU gain of 6.6 to 12.1 compared to multitask learning approaches but can still improve model performance	Data Augmentation achieved a weakly supervised training data to enhance data scarcity but cannot be solely utilized for self-supervised training.
Lee et al., 2022a	Direct Speech-to-Speech Translation with Discrete Units	Apply self-Supervised speech encoder to convert target speech to discrete units	S2UT Model, HuBERT Model and HiFi GAN	The resulting S2UT model achieves only a 1.4 BLEU difference on transformer Translatotron	used synthetic Data. It shows that discrete units are easier to model than mel-spectrogram features
Ning Yang, 2024	Optimization and Performance Analysis	Using CTC for model criterion and parameter compression.	BiLSTM-CTC architecture, for its ability to handle variable-	The introduction of smoothing gates reduced the model's	Even with high parameter reduction, it did led to a

Author(s) and Year	Paper title	Methodology	Algorithm	Results	Remark
	of Real-time Speech Translation Systems Based on Mobile Technology	Dynamic Pruning: A smoothing gate mechanism was introduced to evaluate the importance of memory units within the BiLSTM network. Used best-path decoding for results and prefix-search decoding to refine the final output.	length input sequences and capture contextual information. CTC Criterion addresses the alignment issue between long speech feature sequences and shorter text annotations with Forward-Backward Algorithm integrated to it.	parameters by approximately 42% (from 10.89 M to 6.36 M) and multiplication operations by about 42%. the Word Error Rate (WER) showed minimal change, moving from 14.25% in the baseline to 14.22% in the optimized model.	slight performance drop in WER compared to the baseline. The proposed methods mitigate resource issues; large-scale models still face potential memory shortages and performance bottlenecks on mobile devices compared to server-grade equipment
Sina Rashidi and Hossein Sameti, 2025	Improving direct Persian–English speech-to-speech Translation with discrete units and synthetic Parallel data	built a Direct S2ST system to overcome the limitations of cascaded. Used GPT-3.5 to translate Persian speech transcriptions into English, and then converts that text into English speech using a Text-to-Speech (TTS) system (VoiceCraft). Used HuBERT Model to predicts discrete speech units.	GPT-3.5, S2UT Model, HuBERT Model and HiFi GAN, VoiceCraft	The direct S2ST model achieved a BLEU score of 18.4 on the test set which has lower latency, it slightly underperformed compared to traditional cascaded systems. Initializing the encoder with Wav2Vec 2.0 significantly accelerated convergence and improved the final translation accuracy compared to training from scratch.	Any errors in the GPT-generated text or robotic tones in the synthetic audio are learned by the S2ST model. Recording single-speaker struggled with noisy environments and diverse accents in Persian.
Marie Maltais et al., 2026	NaijaS2ST: A Multi-Accent Benchmark for Speech-to-	The study involved a systematic benchmark	Evaluated three approaches: Cascaded Methods using	Audio LLMs (Gemini 3.1) surpasses both fine-tuned E2E	SeamlessM4T only supports translation into high-resource

Author(s) and Year	Paper title	Methodology	Algorithm	Results	Remark
	Speech Translation in Low-Resource Nigerian Languages	comparing three modeling paradigms: cascaded, end-to-end (E2E), and AudioLLM-based approaches across bidirectional translation settings.	Omnilingual-ASR, NLLB-200 3.3B or Tiny-Aya-Global 3B, Gemini 2.5 Flash TTS. End-to-End (E2E) Methods using Evaluated SeamlessM4T-Large V2 (2.3B), (S2TT) and SeamlessM4T-Large (2.3B). using AudioLLMs: Gemini 2.5, Gemini 3.1, and GPT-Audio 1.5.	models and cascaded baselines. E2E and Audio LLM approaches outperformed traditional cascaded pipelines, due to error propagation from the ASR stage to the translation stage. Fine-tuning significantly improved performance for Hausa; however, while Mono-FT was best for translating into English,	languages. For budgetary reasons, quality-driven re-recording was prioritized for the development and test sets but was not conducted for the training set. Using ChrF were found to be generally unreliable for languages like Yorùbá that utilize extensive diacritics.

In summary, This research draws from Abbott & Martinus (Transformer baseline), Jia et al. (voice preservation challenges and solutions), the textless discrete-unit paradigm (Li, Lee), self-supervised pre-training (Popuri), mobile-friendly compression (Ning Yang), synthetic data generation (Rashidi & Sameti) for augmentation, and the Nigerian - specific benchmarking of Maltais et al. Together, these studies point toward a S2ST system that is optimized by Adam and finetuned using DE for real-time. Limited or no previous study has applied Differential Evolution to fine tune a HuBERT based S2ST model, nor has a direct Igbo to English S2ST system been reported. This work fills that gap.

3. METHODOLOGY

The methodology is to provide a clear, systematic, and structured approach to which this research will be based on.

3.1. Data Acquisition and Preprocessing

A parallel Igbo–English speech corpus was generated by recording fluent bilingual speakers (comprising of female and male) in a quiet environment using a high-quality microphone. The recordings consist of 300 sentence pairs drawn from common educational phrases (e.g., classroom instructions, questions, and feedback) as shown in table 2. Each sentence was recorded separately, yielding 300 Igbo audio files and 300 English audio files. The corpus was split into training (70%, 210 pairs), validation (15%, 45 pairs), and test (15%, 45 pairs).

Table 2: Sample Sentences Used in Training and Testing of the Models

Igbo	English
Azịza gị niile ga-abụ na Bekee.	All your answers must be in English.
Ị nwere ajụjụ ọ bụla?	Are there any questions?
A na-enye anyị ikike iji kọmpụta?	Are we allowed to use calculators?
Ònye nwere ike ịza nke a?	Can anyone answer this?
Ònye nwere ike ịgwa m ihe anyị kpaịrị oge gara aga?	Can anyone tell me what we discussed last time?
Enwere m ike ịza ajụjụ ahụ?	Can I answer the question?
Enwere m ike ịjụ ajụjụ?	Can I ask a question?

All audio files were resampled to 16 kHz using *librosa.resample* and trimmed to remove leading/trailing silence (energy threshold 1% of maximum). For the source (Igbo) side, 80 - dimensional log-mel filterbanks were extracted with a 25 ms window and 10 ms hop length, followed by utterance-level cepstral mean and variance normalization (CMVN). During training, SpecAugment (time and frequency masking) was applied on-the-fly.

For the target (English) side, features were extracted from layer 6 of a pre-trained HuBERT Base model (12 transformer layers, 768 dimensions). These continuous features were quantized into discrete unit sequences using a k-means model with 100 clusters, pre-trained on a large English speech corpus. The resulting integer sequences (e.g., "71 71 12 72 …") served as the decoder's target labels.

3.2. Model Architectural Framework

The research frame work was based on careful consideration of the strengths and suitability of these models for Natural Language Processing. The system design followed a speech-to-unit-to-speech (S2UT) architecture from the Fairseq framework where the input Igbo speech is converted to a sequence of discrete acoustic units using a Hidden Unit Bidirectional Encoder Representation from Transformers (HuBERT) Based sequence-to-sequence model with a speech encoder and a discrete unit decoder. These discrete acoustic units are then synthesized into English speech using a neural vocoder (HiFi-GAN vocoder). The HuBERT Model was trained to facilitate model learning by its flexibility to introduce more parameters to generate more features or relations between the data. It was optimized using the Adam/Differential Evolution (DE) Optimizer.

The encoder consists of a convolutional subsampler (two Conv1d layers with stride 2, kernel size 5, 1024 channels) followed by a 12-layer Transformer encoder (512-dim, 8 heads, 2048 FFN, pre-normalization). The decoder is a 6-layer Transformer (512-dim, 8 heads) with cross-attention over the encoder outputs. The output vocabulary size is 104 (100-unit classes + 4 special tokens). Dropout was set to 0.1 for all layers.

During inference, the model generates a sequence of discrete unit indices (beam size = 5). These units are then fed to a pre-trained HiFi-GAN vocoder (Kong et al., 2020) to synthesize the final English waveform. The research framework is shown in figure 1.

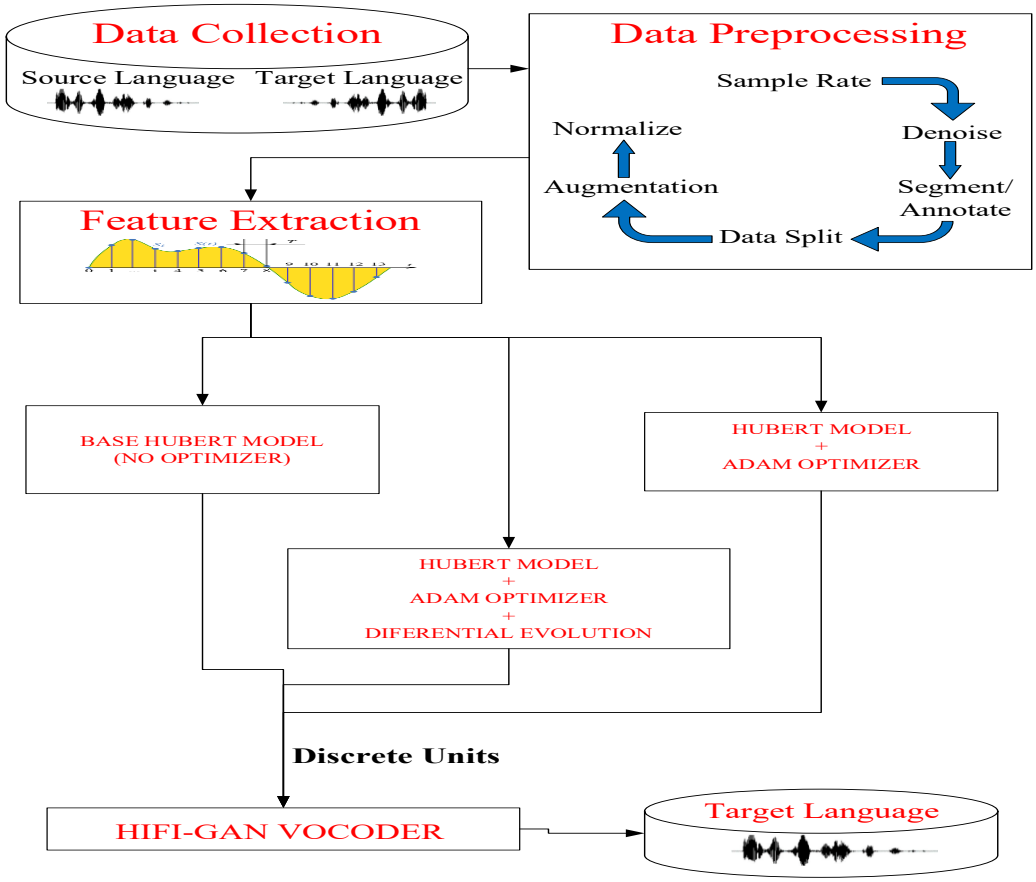


Figure 1: Framework of the Direct speech to speech Model with discrete Units

3.3. Hidden-Unit Bidirectional Encoder Representation from Transformers (HUBERT)

HuBERT: A self-supervised speech representation model developed by Facebook AI Research, employed both for extracting continuous features (from Igbo) and for quantizing English audio into discrete units. It leverages the strengths of both BERT-like architectures and unsupervised learning techniques to represent speech by predicting masked parts of audio input, similar to how BERT predicts masked words in text. It uses a clustering algorithm to create discrete representations (hidden units) of the audio and trains to predict these hidden units over masked audio segments. The HuBERT model is pretrained in a self-supervised manner using a masked prediction task, which works as follows:

Quantization Step: First, HuBERT uses a quantizer to convert the continuous features (like Mel-spectrograms) into discrete units. This is done by clustering the features using a method like k-means to represent speech signals as discrete tokens. The features are mapped to the nearest cluster center. These tokens represent the "hidden units" or "codebook" that will be learned in the next steps. The formular is given as thus:

$$\hat{y} = \arg\min_{k} \|x - \mu_k\|^2 \quad (1)$$

Where $\hat{y}$ is the quantized representation of the input, μ_k is the cluster center, and x is the input feature.

Masking: Random sections of the discrete token sequence are masked, and the loss is computed based on the prediction of the masked tokens similar to how BERT masks tokens in a sentence. Masking is typically done with a binary mask (1,0).

Prediction: The model is trained to predict the masked tokens based on the surrounding (unmasked) context. This involves calculating the hidden states of the model (using a transformer architecture), which can be thought of as the equivalent of "hidden units" in BERT. The transformer-based model

computes the hidden representations using self-attention. For each token, the attention mechanism calculates:

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Where Q is the query, K is the key, and V is the value. This allows the model to weigh the importance of different parts of the sequence and aggregate information across the sequence.

Loss function: A cross-entropy loss function is used to minimize the difference between the predicted tokens sequence and the actual (unmasked) tokens sequence:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (3)$$

Where y_i is the true token and $\hat{y}_i$ is the predicted token

3.4. The Differential Evolution (DE) optimizer

This is a powerful, population-based optimization algorithm that belongs to the family of evolutionary algorithms, similar to genetic algorithms used to solve complex optimization problems, especially in situations where the objective function is nonlinear, non-differentiable, and highly multimodal. The algorithm was introduced by Storn and Price in 1996. it can be formulated as the task of finding a vector of variables $X = x_1\ x_2\ x_3, \ldots., x_n$ inside the search space otherwise known as the population N, to minimize the objective function f(X). Therefore, no any other point in the search space will have an objective value larger than f(X). DE operation consists of three common basic evolutionary steps: mutation, crossover (or recombination) and selection, although its operation has to be initialized and at the end a repeat is activated.

Initialization: A population of N individuals is initialized randomly within the problem's search space. Suppose we want to optimize a function with D real parameters; define upper and lower bounds for each parameter. The parameter vectors have the form:

$$X_{i,G} = x_{1,i,G}\ x_{2,i,G}\ x_{3,i,G}, \ldots., x_{D,i,G} \quad (4)$$

Where i = 1, 2, . . ., N. and G is the generation number.

Mutation: For each individual (parent) in the population, a mutant vector is generated using a mutation strategy. The typical mutation strategy is:

$$V_{i,G+1} = x_{r1,G} + f.(x_{r2,G} - x_{r3,G}) \quad (5)$$

Where $x_{r1,G}$, $x_{r2,G}$, and $x_{r3,G}$ are randomly selected individuals from the population

f is the mutation factor or the scaling factor that controls the mutation.

$V_{i,G+1}$ is the mutant vector or the Donor Vector

Crossover: The mutant vector $V_{i,G+1}$ is mixed with the current individual $X_{i,G}$ to create a trial vector $U_{i,G+1}$. Crossover involves combining corresponding elements from the parent and mutant vectors based on a crossover rate CR, which determines the likelihood that an element will be taken from the mutant vector.

Selection: The trial vector $U_{i,G+1}$ is evaluated using the objective function. If the trial vector improves the objective function, it replaces the current individual $X_{i,G}$. If not, the current individual remains in the population.

Mutation, recombination and selection continue until some stopping criterion is reached, either a pre-defined number of generations is met or until convergence occurs.

3.4. Optimization Strategies

Comparison was carried out based on three optimization configurations:

i. **Hubert Model with No optimizer**: baseline with learning rate effectively zero (no weight updates).

ii. **Hubert Model with Adam Optimizer**: primary training with learning rate 5×10^{-4}. Training ran for up to 100 epochs or 20,000 updates.
iii. **Hubert Model with Adam/Differential Evolution (DE) fine-tuning**: After Adam training, we applied DE (SciPy implementation) to a subset of 300 weights from the decoder's output projection layer ($104 \times 512 = 53{,}248$ total). The search space was bounded to ±0.1 around the original Adam values. Population size was 900 (3×dimension), generations 20.

3.5. Training Dynamics

Figure 2 and figure 3 shows the training and validation loss/Perplexity curves for the No optimizer and Adam trained model respectively. The No Optimizer curve remained steady as no learning was taking place. For the Adam Optimizer, Training loss decreased steadily from 5.76 to 1.52, while validation loss reached a minimum of 1.62 around epoch 97, after which it slightly increased, indicating mild overfitting. The gap between training and validation remained small, confirming good generalization. The DE fine tuning converged after 20 generations, reducing the dev set loss from 1.5189250 to 1.5188616, a negligible change that did not affect the discrete unit predictions as shown in figure 4.

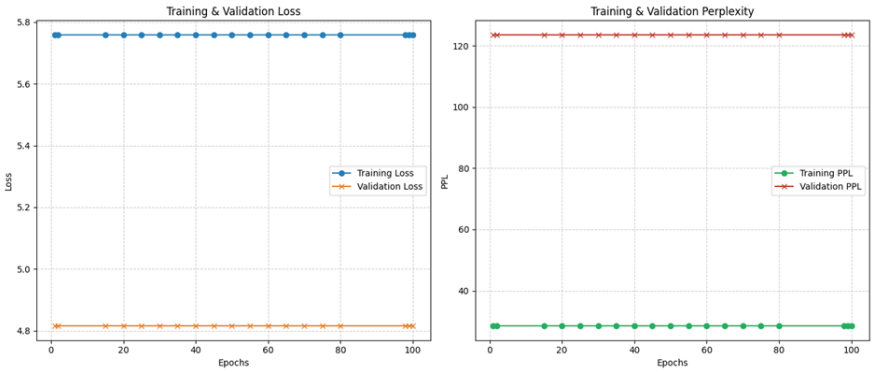


Figure 2: Training and Validation loss/Perplexity trends for HUBERT with No Optimizer

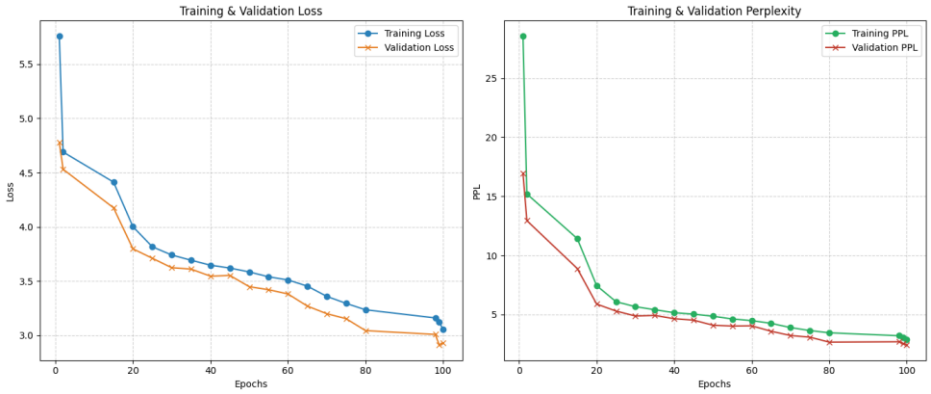


Figure 3: Training and Validation loss/Perplexity trends for HUBERT with Adam Optimizer

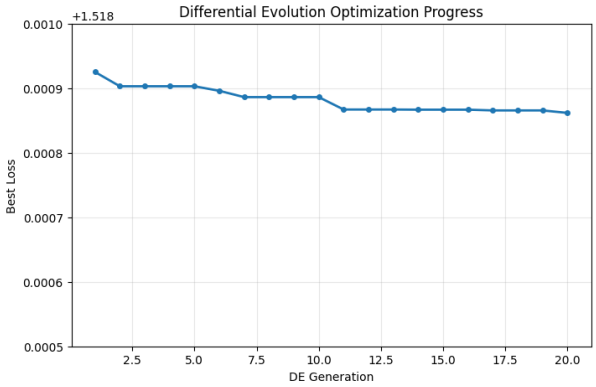


Figure 4: training loss trends for the HUBERT with Adam/DE fine-tuning

3.6. Evaluation Metrics

Evaluation of the translation quality was carried out using these three metrics:

i. **Unit Bilingual Evaluation Understudy (BLEU4):** This metric measures the similarity between the model's generated translations and reference translations based on n-gram overlap. Usually, BLEU is calculated on English words but since this command is run on .unit files, the BLEU4 score is measuring Unit Overlap. The higher the better and the formular is given thus;

$$\text{BLEU} = \exp\left(\min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n\right) \quad (6)$$

ii. **Unit Error Rate (UER)**: This is analogous to Word Error Rate (WER) but applied to unit sequences. It is expressed in percentage and the lower the better. Its formular is expressed as thus:

$$\text{UER} = \frac{\text{S} + \text{I} + \text{D}}{\text{N}} \times 100 \quad (7)$$

iii. **Token-level Accuracy:** Token-level accuracy measures the percentage of correctly predicted unit tokens generated by the model on the held-out dataset, ignoring padding and special tokens (e.g., <eos>). The translations which will be in discrete units can be compared with the target language discrete unit of the Test dataset known as the Ground Truth.

$$\text{Accuracy} = \frac{\text{Number of Correctly predicted tokens}}{\text{Total Number of Tokens in reference}} \times 100 \quad (8)$$

where: “Correctly predicted tokens" are tokens that exactly match the reference

4. RESULTS AND DISCUSSION

The experimental results confirm the system's efficacy and dependability in translating educational content among English and Igbo. A comparative analysis against Hubert baseline models without optimizer, optimized using Adam optimizer and an additional optimization layer using Differential Evolution (DE), applied to a subset of the decoder's output projection weights (300 parameters) with bounds of ±0.1 over 20 generations. The discussion concludes with insights into the significant role of hyperparameter tuning and the inherent advantages of the optimization.

4.1. Optimization Comparison

Table 3 summarizes the performance of the three optimization strategies on the test set. The model without any optimizer failed to learn (BLEU4 = 0.15, UER = 253.14%, accuracy = 1.60%). The Adam-trained model achieved a BLEU4 of 23.97, UER of 66.18%, and accuracy of 33.69%. DE fine-tuning produced identical scores, indicating that Adam had already reached a strong local optimum and the constrained evolutionary search could not improve upon it.

Table 3: Performance comparison of optimization strategies with training convergence time

Methods	Igbo ~ English			
	BLEU4	**UER**	**Accuracy**	**Convergence time**
Hubert + No Optimizer	0.15	253.14%	1.60%	64,800
Hubert + Adam Opt.	23.97	66.18%	33.69%	65,357
Hubert + Adam + DE Opt.	23.97	66.18%	33.69%	29,781 (DE only)

Figure 5 visually corroborates these findings, presenting a comparison of the baseline Adam-trained S2UT Transformer model and the model further optimized using Differential Evolution (DE).

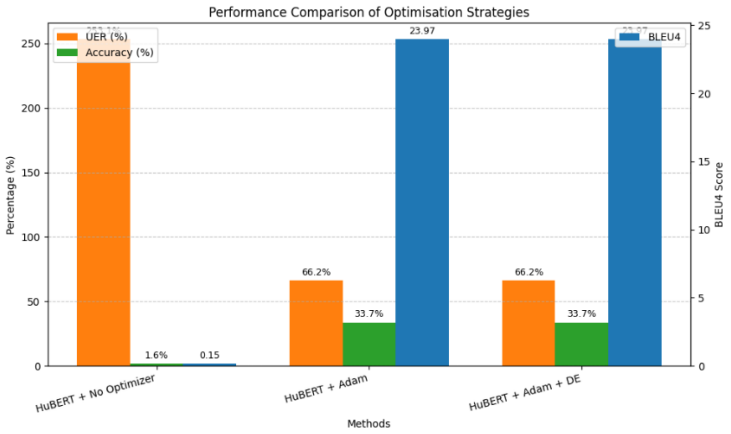


Figure 5: Model performance of the optimized models

The BLEU4 score of 23.97 is reasonable for a low-resource S2ST task. For comparison, Rashidi & Sameti (2025) reported a BLEU score of 18.4 for Persian–English using synthetic data. Our unit-based BLEU is not directly comparable to text-based BLEU because discrete units are more repetitive, but it confirms that the model successfully learned to map Igbo speech to meaningful English acoustic units.

5. CONCLUSION

This paper presented a direct speech-to-speech translation system from Igbo to English using a speech-to-unit (S2UT) framework with HuBERT discrete units. The Adam optimizer trained model achieved a BLEU4 score of 23.97 on a self-collected parallel corpus of 300 utterance pairs. A comparative experiment with Differential Evolution fine-tuning on a subset of the decoder's output projection weights yielded no improvement, confirming that gradient-based optimization already reaches a strong local optimum for this task. The system is deployed as a user-friendly desktop python application, enabling practical Igbo–English translation in educational contexts.

Future work includes: (i) expanding the dataset to at least 1,000 utterance pairs to improve generalization, (ii) exploring advanced evolutionary algorithms (e.g., CMA-ES) on larger parameter subsets, (iii) optimizing the model for real-time edge deployment (mobile phones) using pruning and quantization, and (iv) extending the system to support additional low-resource Nigerian languages (e.g., Hausa, Yorùbá) through transfer learning.

REFERENCES

Abbott, J., & Martinus, L. (2018). Towards neural machine translation for African languages. *Proceedings of the 2nd Workshop on African Natural Language Processing*. https://doi.org/10.48550/arXiv.1811.05467

Abiola, O. B., Adetunmbi, A. O., & Oguntimilehin, A. (2015). Using hybrid approach for English-to-Yoruba text to text machine translation system. *International Journal of Computer Science and Mobile Computing*, *4*(8), 308–313.

Dossou, B. F. P., & Emezue, C. C. (2021). *OkwuGbé: End-to-end speech recognition for Fon and Igbo*. arXiv. https://doi.org/10.48550/arXiv.2103.07762

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units*. arXiv. https://doi.org/10.48550/arXiv.2106.07447

Jia, Y., Ramanovich, M. T., Remez, T., & Pomerantz, R. (2022). Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *Proceedings of the 39th International Conference on Machine Learning*, 10120–10134. https://doi.org/10.48550/arXiv.2107.08661

Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., & Wu, Y. (2019). Direct speech-to-speech translation with a sequence-to-sequence model. *Proceedings of Interspeech 2019*, 1123–1127. https://doi.org/10.21437/Interspeech.2019-1951

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high-fidelity speech synthesis. *Advances in Neural Information Processing Systems*, *33*, 17022–17033. https://doi.org/10.48550/arXiv.2010.05646

Lee, A., Chen, P.-J., Wang, C., Gu, J., Popuri, S., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., & Hsu, W.-N. (2022a). Direct speech-to-speech translation with discrete units. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3327–3339. https://doi.org/10.18653/v1/2022.acl-long.235

Lee, A., Gong, H., Duquenne, P.-A., Schwenk, H., Chen, P.-J., Wang, C., Popuri, S., Adi, Y., Pino, J., Gu, J., & Hsu, W.-N. (2022b). Textless speech-to-speech translation on real data. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 860–872. https://doi.org/10.18653/v1/2022.naacl-main.63

Li, X., Jia, Y., & Chiu, C.-C. (2022). Textless direct speech-to-speech translation with discrete speech representation. *arXiv preprint arXiv:2211.00115*. https://doi.org/10.48550/arXiv.2211.00115

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, *40*(3), Article 8. https://doi.org/10.1145/1380584.1380586

Maltais, M., et al. (2026). *NaijaS2ST: A multi-accent benchmark for speech-to-speech translation in low-resource Nigerian languages*. arXiv. https://doi.org/10.48550/arXiv.2604.16287

Nwonu, E. G., Nkamaigbo, L. C., & Ezenwafor-Afuecheta, C. I. (2023). Multilingualism and insecurity in Nigeria. *International Journal of Advanced Academic Studies*, *10*(2), 45–58.

Omachonu, G. S. (2012, October 20). *Igala language studies and development: Progress, issues and challenges* [Paper presentation]. 12th Igala Education Summit, Kogi State University, Anyigba, Nigeria.

Popuri, S., Chen, P.-J., Wang, C., Pino, J., Adi, Y., Gu, J., Hsu, W.-N., & Lee, A. (2022). Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *Proceedings of Interspeech 2022*, 5195–5199. https://doi.org/10.21437/Interspeech.2022-11032

Rashidi, S., & Sameti, H. (2025). *Improving direct Persian–English speech-to-speech translation with discrete units and synthetic parallel data*. arXiv. https://doi.org/10.48550/arXiv.2511.12690

Storn, R., & Price, K. (1997). Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762