Multimodal Translation Pipelines for Low-Resource African Languages

Adrian Castillo
CS 697R

January 6, 2026

Abstract

This project evaluates ten speech-to-speech translation pipelines for four low-resource African languages: Efik, Igbo, Swahili, and Xhosa. We fine-tuned NLLB-600M for machine translation and XTTS models for text-to-speech synthesis. The pipelines were evaluated using BLEU, chrF, and COMET for translation quality, and MCD and BLASER 2.0 [3] for audio quality.

1 Introduction

This project explores speech and text translation pipelines for low-resource African languages. The goal is to evaluate different pipeline architectures that combine neural machine translation and text-to-speech synthesis for four African languages: Efik, Igbo, Swahili, and Xhosa.

2 Methodology

2.1 Datasets

The datasets for all four African languages—Efik, Igbo, Swahili, and Xhosa—were provided by the Pathsay program. Synthetic English audio was generated from the source English text of the dataset to be used as the input for audio-to-audio pipeline evaluations. The model used for generation was Whisper (openai/whisper-medium).

2.2 Models

The model used for machine translation was NLLB-600M [1]. For text-to-speech, we fine-tuned six XTTS [2] model variants. Some of these variants receive as input text in the corresponding output language, while for others we experimented with using English text as input and the target audio in the African language. The XTTS variants are as follows:

Model Name	Input	Output
Native	African text	African audio
Eng2Multi	English text	African audio
Eng2Efik	English text	Efik audio
Eng2Swa	English text	Swahili audio
BiTag	`<eng>{eng} <lang>{lang}`	African audio
TransTag	`<translate> <eng>{eng} <lang>{lang}`	African audio

2.3 Pipeline Architectures

Using NLLB and the TTS variants, several translation pipelines were constructed. Since some of the TTS variants used the raw English text as input, we treat them as audio-to-audio pipelines.

Pipeline	Translation	TTS Model
Pipeline 1	NLLB	Native
Pipeline 2	NLLB	BiTag
Pipeline 3	None	BiTag
Pipeline 4	NLLB (custom format)	BiTag
Pipeline 5	NLLB	TransTag
Pipeline 6	None	TransTag
Pipeline 7	NLLB (custom format)	TransTag
Pipeline 8	None	Eng2Multi
Pipeline 9	None	Eng2Efik (Efik only)
Pipeline 10	None	Eng2Swa (Swahili only)

2.4 Evaluation Metrics

The text metrics used for NLLB are BLEU, chrF, and COMET (McGill-NLP/ssa-comet-qe). For audio metrics, we used Mel-Cepstral Distance (MCD) and BLASER 2.0 (facebook/blaser-2.0-qe). MCD measures spectral distance between predicted and reference audio; lower values are better. BLASER 2.0 evaluates semantic similarity between source and predicted audio, with scores ranging from 1 to 5. BLASER does not support Efik, Igbo, and Xhosa out of the box; speech encoders were fine-tuned (including Swahili) for evaluation purposes.

2.5 Framework

The fine-tuned NLLB model generated translations for 300 samples. Using these samples, we either took the translation or the English source text to generate audio using the XTTS models. Once the audio was generated, we evaluated it using the audio metrics.

3 Results

3.1 Translation Quality

Table 1 shows NLLB translation quality for the samples used. Swahili achieves the highest BLEU (49.2), chrF (76.0), and COMET (0.676) scores. Efik has the lowest scores (BLEU: 29.8), which is expected since it was the only language not originally supported by NLLB.

Language	BLEU	chrF	COMET
Efik	29.842 ± 18.580	58.958 ± 14.244	0.603 ± 0.068
Igbo	32.953 ± 23.539	60.635 ± 17.772	0.641 ± 0.068
Swahili	49.217 ± 18.507	76.009 ± 10.479	0.676 ± 0.051
Xhosa	32.422 ± 20.713	71.084 ± 12.767	0.649 ± 0.048
Overall	36.109 ± 7.659	66.671 ± 7.117	0.642 ± 0.026

Table 1: Translation Quality Metrics for NLLB. Values are Mean ± Std. Higher is better for all metrics.

3.2 Audio Quality

Table 2 shows MCD and BLASER scores across all pipelines.

Metric	Pipeline	Efik	Igbo	Swahili	Xhosa	Overall
MCD	NLLB → Native	**13.123 ± 1.348**	**12.956 ± 1.212**	**13.398 ± 0.950**	**11.887 ± 1.058**	**12.841 ± 0.573**
	NLLB → BiTag	13.334 ± 1.365	13.297 ± 1.212	13.793 ± 0.903	12.474 ± 0.972	13.224 ± 0.475
	Source → BiTag	13.814 ± 1.299	13.968 ± 1.141	14.256 ± 0.813	13.042 ± 0.808	13.770 ± 0.449
	Custom Lang → BiTag	13.169 ± 1.392	13.051 ± 1.177	13.407 ± 0.946	11.969 ± 1.084	12.899 ± 0.552
	NLLB → TransTag	13.331 ± 1.328	13.445 ± 1.262	13.878 ± 0.893	12.482 ± 1.037	13.284 ± 0.506
	Source → TransTag	13.776 ± 1.278	13.947 ± 1.213	14.271 ± 0.832	12.991 ± 0.830	13.746 ± 0.471
	Custom Translate → TransTag	13.171 ± 1.339	13.088 ± 1.199	13.444 ± 0.986	11.901 ± 1.098	12.901 ± 0.592
	Source → Eng2Multi	13.643 ± 1.200	13.870 ± 1.132	14.043 ± 0.832	12.875 ± 0.825	13.608 ± 0.446
	XTTS Eng2Efik	13.716 ± 1.189	—	—	—	—
	XTTS Eng2Swa	—	—	13.949 ± 0.789	—	—
BLASER	NLLB → Native	2.728 ± 0.221	3.087 ± 0.239	2.585 ± 0.281	2.600 ± 0.292	2.750 ± 0.202
	NLLB → BiTag	**2.750 ± 0.218**	**3.090 ± 0.254**	2.561 ± 0.281	2.763 ± 0.290	2.791 ± 0.190
	Source → BiTag	2.682 ± 0.213	2.906 ± 0.282	2.756 ± 0.261	2.778 ± 0.273	2.781 ± 0.081
	Custom Lang → BiTag	2.723 ± 0.221	3.086 ± 0.242	2.574 ± 0.281	2.599 ± 0.293	2.746 ± 0.204
	NLLB → TransTag	2.749 ± 0.217	3.086 ± 0.254	2.568 ± 0.284	2.748 ± 0.289	2.788 ± 0.187
	Source → TransTag	2.708 ± 0.209	2.983 ± 0.267	**2.851 ± 0.238**	**2.879 ± 0.270**	2.855 ± 0.098
	Custom Translate → TransTag	2.725 ± 0.224	3.088 ± 0.241	2.575 ± 0.276	2.597 ± 0.302	2.746 ± 0.205
	Source → Eng2Multi	2.722 ± 0.214	3.074 ± 0.249	2.844 ± 0.236	2.850 ± 0.236	**2.873 ± 0.127**
	XTTS Eng2Efik	2.715 ± 0.207	—	—	—	—
	XTTS Eng2Swa	—	—	2.813 ± 0.244	—	—

Table 2: Audio and Speech Quality Metrics Across All Pipelines. Values are Mean ± Std. **Bold** indicates best performance per language. MCD: lower is better; BLASER: higher is better.

For MCD, Pipeline 1 (NLLB → Native) achieves the best scores across all languages with an overall MCD of 12.841. Xhosa has the lowest MCD (11.887) in Pipeline 1.

For BLASER, Pipeline 8 (Source → Eng2Multi) achieves the best overall score (2.873). Pipeline 2 (NLLB → BiTag) has the highest BLASER scores for Efik (2.750) and Igbo (3.090). Pipeline 6 (Source → TransTag) has the highest scores for Swahili (2.851) and Xhosa (2.879).

4 Discussion

The results showed that BLASER scores across all pipelines were very similar. This raises some questions: it could be related to the fact that the same samples were used for all evaluations, or it might be due to the way the speech encoders were fine-tuned. It's also possible that some part of the evaluation did not work as intended, though this would require further investigation.

An additional evaluation, not included in the main results, used the target text from the Pathsay datasets directly. For Swahili, the audio quality in this evaluation reached scores above 4. However, when using translations generated by NLLB, the scores decreased to around 2–3. This suggests that the quality of machine translation directly impacts the audio quality, and future work should explore ways to improve translation accuracy and study its effect on the final audio output.

On a practical note, this project provided valuable hands-on experience. I learned to work with a supercomputer in multiple scenarios, processed large amounts of data, fine-tuned several models, and built an application to run the pipelines. While I spent more time than expected on the application, it gave me insights into how ML models interact with servers for inference. The application followed a microservice architecture with an API gateway orchestrating audio generation, allowing actual translations using the trained models. This experience helped me understand the end-to-end workflow of deploying ML systems.

5 Conclusion

This project evaluated ten speech-to-speech translation pipelines for four low-resource African languages. Swahili consistently achieved the highest translation and audio quality, while Efik was the most challenging due to limited support in NLLB. BLASER scores were very similar across pipelines, possibly due to shared evaluation samples or the fine-tuning of speech encoders.

The additional evaluation using target text highlighted that audio quality is strongly influenced by translation accuracy. This underscores the importance of improving machine translation for low-resource

languages to achieve better end-to-end speech translation.

Overall, the results demonstrate that combining NLLB with fine-tuned XTTS models can produce intelligible and semantically accurate speech translations. Future work should focus on improving translation quality, refining evaluation methods, and exploring more robust pipeline architectures for low-resource languages.

References

[1] NLLB Team et al., *No Language Left Behind: Scaling Human-Centered Machine Translation*, arXiv preprint arXiv:2207.04672, 2022.

[2] Casanova, E., et al., *XTTS: Massively Multilingual Zero-Shot Text-to-Speech*, Coqui.ai Technical Report, 2023.

[3] David Dale and Marta R. Costa-jussà, *BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation*, Findings of the Association for Computational Linguistics: EMNLP 2024, 2024.