Transfer Learning for Creole TTS: A Pilot Study on Whether Substrate Phonologies or Lexifier Vocabularies Matter More

Emmett Strickland[1], Marc Evrard[2], Valentina Fedchenko[1]
[1]National Institute for Oriental Languages and Civilizations (INALCO), France
[2]Paris-Saclay University, France
emmett.strickland@inalco.fr, marc.evrard@universite-paris-saclay.fr, valentina.fedchenko@inalco.fr

Abstract

In this early-stage study, we investigate whether transfer learning from lexifier or substrate languages can improve text-to-speech (TTS) performance for low-resource creoles. We conducted a controlled experiment using two creoles of distinct lexical origins: Nigerian Pidgin (English-based) and Guadeloupean Creole (French-based). Single-speaker TTS datasets of approximately 30 minutes each were recorded and used to fine-tune pretrained models for English, French, and Yoruba. Objective metrics and informal subjective evaluations were employed to assess synthesis quality. Though partially inconclusive, our results suggest that the French-based models outperform others for both creoles, while Yoruba-based models yield weaker performance. These findings may suggest that lexical similarity or historic influences alone do not fully predict transfer learning effectiveness, and that phonotactic compatibility and orthographic depth may also be relevant factors. Our work provides insight into TTS model development for creoles and other low-resource languages, and highlights avenues for further research on leveraging relevant linguistic and orthographic features for model development.

Keywords: creoles, low-resource languages, transfer learning, text-to-speech (TTS)

1. Introduction

1.1. Transfer learning for low-resource TTS and NLP tasks

Neural text-to-speech (TTS) architectures allow researchers to create robust, from-the-ground-up speech synthesis models with just a few hours of training data. However, transfer learning with pretrained models can be an effective alternative when data is scarce, as with low-resource languages without dedicated NLP datasets (Soria et al., 2013). By finetuning a model trained on a better-resourced relative, one can exploit structural similarities with the target language and build a model with just 30 minutes of additional data. For example, Orjuela et al. (2025) managed to build a functional TTS model for the endangered Romance language Walloon by finetuning a pretrained French model with just 18 minutes of speech.

1.2. Creoles and why they are interesting for transfer learning

The notion of 'relatives' is easy to define when it comes to languages that evolved from a recent common ancestor. However, creole languages represent a more complex case. While broadly described as 'mixed' languages, the exact mechanisms by which these varieties emerge and how their parent languages contribute to them are hotly debated.

About 100 languages commonly described as 'creoles' are spoken in the world today. These invariably low-resource languages range from emerging giants like West African Pidgin, to critically endangered varieties like the Gullah-Geechee and Louisiana creoles of the Southern United States. However ubiquitous the term, specialists regularly spar over the fundamental question of what a 'creole' actually *is*. Some scholars consider these to be a typologically distinct class defined by the extreme morphological reductions of their parent languages, a result of the mass interruption of normal intergenerational language transmission seen on the sugar plantations of Jamaica, Haiti, and Suriname during the Atlantic Slave Trade (Bakker et al., 2011; McWhorter, 2001; Bickerton, 1998). Others view creoles primarily as a historical construct, arguing they are shaped by the same processes of change and language contact as 'ordinary' languages like English, French, and Persian (Mufwene, 2020; Aboh, 2016; DeGraff, 2005).

However we define creoles, the languages that shape them can typically be divided into *lexifiers*, the (often European) languages that provide their vocabulary, and *substrates*, the (often African) languages that mainly influence syntax and phonology. For example, many European-lexified West African creoles are nevertheless thought to have developed lexical tone via substrate influences (Yakpo, 2021). Since each input language makes such distinct contributions to the creole's structure, it is worth asking which of the two is most useful for transfer learning tasks. Recent work by Havard et al. (2025) has shown that Haitian automatic speech recognition (ASR) tools fine-tuned on pre-trained

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 101–106
11–12 May 2026.

French models outperform those pre-trained on English. This confirms the utility of lexifiers in transfer learning tasks, but ignores the potential of African substrates.

This is regrettable, as the unique histories of creoles could inform how transfer learning might best be leveraged to work with these languages in a low-resource context. Conversely, experiments in transfer learning could yield new insights as to which languages contribute most to a creole's features, and which of these features matter most for NLP tasks.

1.3. A controlled TTS finetuning experiment

In this paper, we explore these questions more fully in the context of text-to-speech tasks. Namely, we will compare the effectiveness of transfer learning approaches in developing TTS systems for two creoles of different lexical stocks. We will simulate identical low-resource settings for both languages, working with only 30 minutes of equivalent data per creole.

The first is Naijá, an English-lexifier creole also known as Nigerian Pidgin (ISO code `pcm`). Despite the common lexical base with English, Naijá has traditionally been described as a phonologically distinct, syllable-timed tone language (Mafeni, 1971; Faraclas, 1996). The second is Guadeloupean Creole (`gcf`), a French-based creole spoken in the Caribbean. We will build equivalent TTS datasets for each language to fine-tune English and French models pretrained under identical parameters. We hypothesize that lexical similarities will improve performance when the target creole's model is pretrained on its corresponding lexifier.

We also fine-tune a pretrained Yoruba model. As one of West Africa's largest indigenous languages, Yoruba represents an important substrate in Naijá's development and is widely spoken alongside it today. As a Niger-Volta language, it also serves as a pragmatic typological proxy for Fon and other African substrates of Guadeloupean Creole, sharing key features such as an isolating morphology and a dense lexical tone system (Lefebvre and Brousseau, 2011; Bamgbose, 2000). Finally, Yoruba is among the few relevant languages with a sizeable, studio-recorded speech dataset comparable in quantity and quality to our English and French resources.

We hypothesize that the pretrained Yoruba model will benefit Naijá TTS more than English because of phonological similarities.[1] We do not expect this to be especially beneficial for the more geographically far-flung Guadeloupean, whose phonology has had more time to diverge from its African substrates.

[1] In addition to the aforementioned prosodic features, Nigerian Pidgin is notably defined by the loss of English's interdental fricatives and a vowel inventory akin to indigenous languages (Faraclas, 1996).

2. Experiment design

2.1. Data collection

For both Nigerian Pidgin and Guadeloupean Creole, we recorded a dedicated TTS dataset composed of read text. In both cases, we recorded a single female native speaker reading about 30 minutes of prose in a studio setting. Both volunteers were asked to re-record sentences in which they laughed, hesitated, or misspoke while reading. For both recording sessions, we used a RØDE Wireless GO II microphone and a sampling rate of 22,050 Hz. We considered the use of custom recordings to be essential for mitigating differences in audio quality and discourse genre, a common hurdle for languages with limited available data.

We then segmented and aligned the recordings at the sentence level. When possible, sentences lasting longer than ten seconds were subdivided into syntactically cohesive breath groups. Finally, we listened to each segment to ensure it fully corresponded to its transcription and modified the text when appropriate. Both participants provided their explicit written consent for the recordings to be used to build TTS models in an internal research setting.

2.2. Model training

Separately, we trained two single-speaker female models for the creoles' English and French lexifiers. For our English model, we used the LJ Speech dataset of North American English (Ito and Johnson, 2017), while the French model was based on the SIWIS speech dataset of Metropolitan French (Yamagishi et al., 2017). Because the latter contains sentences read with varying degrees of emphasis, we limited our training data to recordings annotated as emphatically 'neutral'. A third female model was also trained from Yoruba using the TTS portion of the ÌròyìnSpeech dataset (Ogunremi et al., 2024).

Each of these datasets was used to train a dedicated English, French, and Yoruba model using the CoquiTTS implementation of the VITS architecture (Eren et al., 2021; Kim et al., 2021). This setup was notably chosen for its compatibility with grapheme-based inputs and the relative simplicity of its preprocessing pipeline, which does not require phoneme alignments for training. To control for the availability of training data, 3000 samples were randomly selected from each dataset to train for 1000 epochs with a batch size of 32. We otherwise used the default CoquiTTS parameters and

grapheme text inputs. All files were resampled to 22,050 Hz prior to training.

Each of these three base models were then fine-tuned for an additional 600 epochs using 300 samples from our Nigerian Pidgin and Guadeloupean Creole corpora, producing a total of six models.

Base Model 3k samples × 1k epochs	**Fine-tuned Models** 300 samples × 600 epochs
English (en)	en_pcm en_gcf
French (fr)	fr_pcm fr_gcf
Yoruba (yo)	yo_pcm yo_gcf

Table 1: Base models and their fine-tuned versions for Nigerian Pidgin (pcm) and Guadeloupean Creole (gcf).

2.3. Evaluation

The remaining Nigerian Pidgin and Guadeloupean Creole recordings, representing about 10% of each dataset, were reserved as a gold standard test set for our objective evaluations.

For each of the six fine-tuned models, we synthesized the target language's evaluation set and computed two measures of distance between the synthesized outputs and the gold standard. Mel Cepstral Distortion (MCD) was measured with the pymcd library[2] as a measure of acoustic similarity. Meanwhile, the Perceptual Evaluation of Speech Quality (PESQ) metric was used to approximate similarity according to human perception. The latter was computed using the pesq Python library[3], with all audio files downsampled to 16,000 Hz to meet the package requirements. A higher MCD score indicates greater acoustic divergence, while a higher PESQ score indicates greater perceptual similarity.

Finally, we conducted an informal subjective evaluation by listening to each model's output and recording our impressions of their intelligibility and naturalness.

3. Results

3.1. Nigerian Pidgin models

Counterintuitively, our French-based model outranked both the English and Yoruba-based models in terms of mean MCD scores generated from our gold standard dataset. As shown in figure 1, models fine-tuned from the English lexifier (17.36) and the Yoruba substrate (17.33) performed similarly,

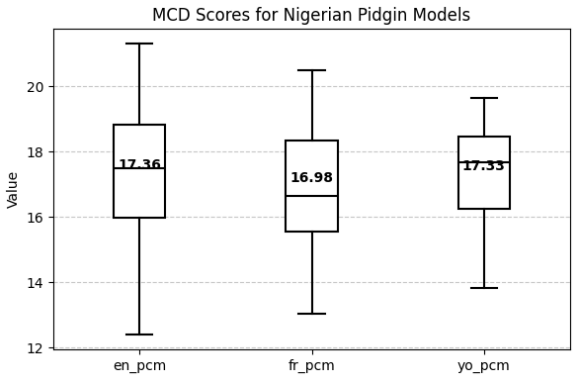

Figure 1: Mel Cepstral Distortion of Nigerian Pidgin models

though the MCD scores produced from the English-based model were characterized by far greater variation. Applying a Friedman test yielded a p-value of 0.22, indicating insufficient evidence to conclude a statistically significant difference between the models.

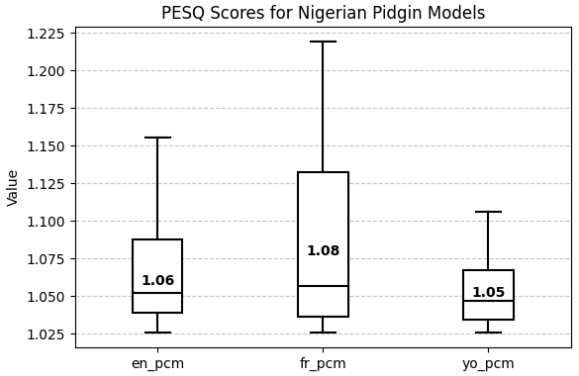

Figure 2: PESQ Scores for Nigerian Pidgin models

The PESQ scores presented in figure 2 show a similar pattern. The French-based model yielded the highest mean perceptual quality score (1.08), with slightly lower results for the models based on English (1.06) and Yoruba (1.05). Despite similar averages, the upper quartiles of French outputs are far higher than those of the other models, indicating a higher quality ceiling. However, a Friedman test yields a p-value of 0.51, meaning that we cannot conclude a statistically significant difference.

During our subjective evaluations, we found that all three models yielded intelligible but noticeably artificial speech. Broadly speaking, the English model appeared to produce polysyllabic words and consonant clusters well, but suffered from buzzing and other noticeable artifacts. We judged the Yoruba model to be acoustically closest to the original speaker in terms of vowel quality. However, the output regularly deteriorated at consonant

[2]https://pypi.org/project/pymcd/
[3]https://pypi.org/project/pesq/

clusters. These often yielded repeated syllables, epenthesized vowels, and unnatural pauses between speech sounds.

The French-based model, though noticeably artificial, handled consonant clusters well and had less frequent buzzing than the English-based model. We found its output to be the most natural on average.

3.2. Guadeloupean Creole models

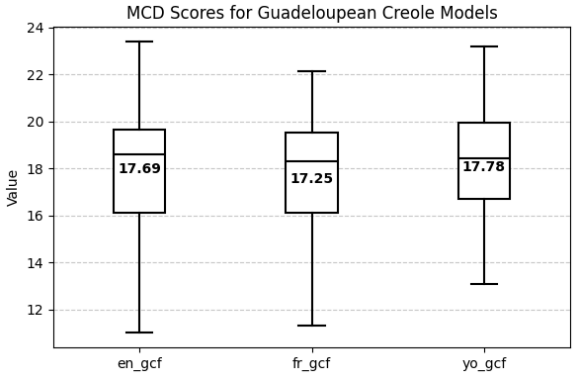

Figure 3: Mel Cepstral Distortion of Guadeloupean Creole models

The MCD scores shown in figure 3 indicate that our French base model also outclassed the others on Guadeloupean Creole, with a mean MCD score of 17.25 compared to 17.69 for English and 17.78 for Yoruba. Once again, the English-based model is characterized by the greatest dispersion of individual scores, while the Yoruba-based model is most consistent. The Friedman test yields a p-value of 0.12, indicating that this difference may still be attributable to chance.

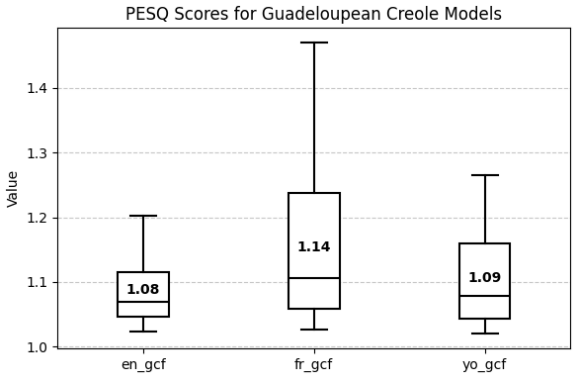

Figure 4: PESQ Scores for Guadeloupean Creole models

The PESQ scores in Figure 4 also show a clear preference for the French-based model, which yielded a higher mean score (1.14) and far greater quality ceiling than the Yoruba and English models. This is also the only instance in which the Yoruba model somewhat exceeded the English model, with a mean PESQ score of 1.08 compared to 1.09. In this case, the Friedman test yields a p-value of 0.0006, indicating a statistically significant difference in model performance.

The MCD and PESQ rankings are largely coherent with our subjective evaluations. The French-based model is by far the most natural and generates the fewest audible artifacts. The English-based model is more prone to occasional buzzing and has somewhat less natural prosody. However, we found the Yoruba model to be significantly weaker, with poorly realized consonant clusters and generally halting speech. Nevertheless, we judged all three models to be noticeably more natural than any of the Naijá models.

4. Discussion

Our results challenge, but do not entirely invalidate our assumption that transfer learning from a creole's parent languages meaningfully improves TTS performance. Both our subjective and objective metrics suggest that French was the most useful language for building a Naijá TTS model, though none of the calculated scores were statistically conclusive. However, the PESQ scores calculated on Guadeloupean Creole did show a statistically significant preference for the French model, which was supported by our subjective impressions. In future research, it will be important to implement formal perceptual tests to see if they validate our findings.

A plausible explanation is that the original French dataset is characterized by some fundamental acoustic difference that made it more compatible with our recordings, and that this factor outweighed any marginal benefit to using the lexifiers or substrate languages. However, it is also worth questioning our definition of 'lexifier' in this context. While Nigerian Pidgin inherited most of its vocabulary from English, many of these items in turn came from French. It is possible that the French model enjoyed the same lexifier benefit as English, with an additional boost from other features it coincidentally shares with Naijá. Namely, these languages share a syllable-timed prosodic structure and frequent vowel nasalization. We also suspect that phonotactic compatibility is an important factor. The fine-tuned Yoruba models consistently struggled to reproduce Naijá and Guadeloupean Creole consonant clusters, a likely consequence of Yoruba's restrictive preference for V and CV syllable structures (Pulleyblank, 2003). Even if the Yoruba-based model generated the most convincing Naijá vowel quality, we found the realization of consonant clusters to be far more important for

perceived quality.

It is also noteworthy that the Guadeloupean Creole models consistently outperformed their Nigerian equivalents. This may be partially attributable to a difference in speaker background and literary practice. Indeed, our Guadeloupean participant was a habitual reader of creole texts, while the Nigerian participant found pidgin prose to be a novelty bordering on delightful.[4] Despite our editing and her satisfactory performance after a warmup exercise, her inexperience with written Naijá gave the recordings a somewhat less fluid quality overall.

It is also worth considering the role of orthographic depth. Our Nigerian Pidgin texts preserve the opaque spellings of English words, while Guadeloupean Creole uses a shallow orthography characterized by close grapheme-to-phoneme mapping. This more regular orthography may have allowed the Guadeloupean models to achieve better performance with the same quantity of training data. Going forward, we hope to reproduce this experiment with phonetized texts to better separate the influences of phonology and orthography. The relationship between spelling and prosody also deserves greater attention going forward. While pitch distinctions are explicitly encoded in Yoruba spelling, this is not the case for Nigerian Pidgin. Any potential benefit from the supposed prosodic similarities between these languages might have been impeded by this fundamental orthographic difference. In future studies, we should also evaluate the base models themselves to see if their performance influences that of their fine-tuned derivatives.

5. Conclusion

We have presented a controlled transfer learning experiment to evaluate whether TTS models for low-resource creoles benefit from finetuning on their lexifier or substrate languages. While our French model clearly benefits our French-based creole, it also appears to help our English-based creole, even if the effect is statistically inconclusive. For both creoles, model performance is weakest when fine-tuned from the African language Yoruba.

Our results are insufficient to conclude that TTS systems for creoles systematically benefit from the languages that shaped them. However, further research with a wider range of languages is warranted. We suspect that historic proximity between languages may not be a central factor in determining the effectiveness of TTS finetuning tasks. Though inconclusive, our results may reassure researchers working with languages without close, well-resourced relatives.

[4] See comments on laughter in section 2.1

6. Acknowledgments

This work is supported by the French National Research Agency and the Ministry of Higher Education, Research and Innovation (MESR). We also extend our gratitude to the Guadeloupean and Nigerian volunteers who lent their voices to this experiment.

7. Bibliographical References

Enoch O Aboh. 2016. Creole distinctiveness: A dead end. *Journal of Pidgin and Creole Languages*, 31(2):400–418.

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole languages*, 26(1):5–42.

Ayọ Bamgbose. 2000. *A grammar of Yoruba*, volume 5. Cambridge University Press.

Derek Bickerton. 1998. Creole languages, the language bioprogram hypothesis, and language acquisition. In *Handbook of child language acquisition*, pages 195–220. Brill.

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in society*, 34(4):533–591.

Gölge Eren, Coqui TTS Team, et al. 2021. Coqui tts. *Zenodo*.

Nicholas Faraclas. 1996. *Nigerian Pidgin*. Routledge, London/New York.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Speech technologies with fieldwork recordings: the case of haitian creole. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 40–46.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pages 5530–5540. PMLR.

Claire Lefebvre and Anne-Marie Brousseau. 2011. *A grammar of Fongbe*, volume 25. Walter de Gruyter.

Bernard Mafeni. 1971. Nigerian pidgin. In John Spencer, editor, *The English language in West Africa*, pages 95–112. Longman, London.

John H McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic typology*, 5.

Salikoko S Mufwene. 2020. Creoles and pidgins: Why the latter are not the ancestors of the former. In *The Routledge handbook of language contact*, pages 300–324. Routledge.

Jose Felipe Espinosa Orjuela, Philippe Boula de Mareüil, and Marc Evrard. 2025. Speech synthesis for walloon, an under-resourced minority language. In *Proc. SSW 2025*, pages 189–195.

Douglas Pulleyblank. 2003. Yoruba. In *The World's Major Languages*, pages 971–990. Routledge.

Claudia Soria, Joseph Mariani, and Carlo Zoli. 2013. Dwarfs sitting on the giants' shoulders–how lts for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, pages 73–79.

Kofi Yakpo. 2021. Creole prosodic systems are areal, not simple. *Frontiers in psychology*, 12:690593.

8. Language Resource References

Keith Ito and Linda Johnson. 2017. *The LJ Speech Dataset*.

Ogunremi, Tolulope and Tubosun, Kola and Aremu, Anuoluwapo and Orife, Iroro and Adelani, David Ifeoluwa. 2024. *ÌròyìnSpeech*. ELRA, ISLRN 012-405-700-001-6.

Yamagishi, Junichi and Honnet, Pierre-Edouard and Garner, Philip and Lazaridis, Alexandros. 2017. *The SIWIS French Speech Synthesis Database*.