Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Yoruba-Speaking Google Maps Navigation

Fiyinfoluwa Oyesanmi
oyesanmifiyinfoluwa@gmail.com

University of Johannesburg

Peter Olukanmi
University of Johannesburg

Research Article

Keywords: Natural Language Processing, Speech-to-Speech Translation, Yoruba Language, Machine Translation, Low

Posted Date: July 23rd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4636009/v1

License: 🅭 🅯 This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Towards Yoruba-Speaking Google Maps Navigation

Fiyinfoluwa Oyesanmi
School of Electrical and Electronics Engineering
University of Johannesburg
Johannesburg, South Africa
0000-0003-1439-1498

Peter Olukanmi
School of Electrical and Electronics Engineering
University of Johannesburg
Johannesburg, South Africa
polukanmi@uj.ac.za

Abstract—Advances in natural language processing (NLP) have made several technological interventions and services available to people in different languages. One such service is the Google Maps direction narration which provides real-time oral assistance to tourists, and visitors in a new or unknown location. Like most related assistive technologies, this service is primarily developed in the English language with support for some other Western languages over time, and the African languages are largely neglected. This paper seeks to leverage advances in NLP techniques and models in the design of a speech-to-speech (STS) translation of the Google Maps direction narration in English to the Yoruba language, one of the most widely spoken languages in Western Africa. We begin with an exploration of various state-of-the-art NLP techniques for Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-speech (TTS) models that make up the designed system. We presented the performance of the models we explored towards the design and implementation of a robust STS translation of the Google Maps direction narration in the Yoruba language.

Index Terms—Natural Language Processing, Speech-to-Speech Translation, Yoruba Language, Machine Translation, Low-Resourced Languages

I. Introduction

The applications of natural language processing, especially in real-time translation have significantly increased over time, offering profound capabilities to make life and living better. One such application is the Google Map Direction narration for immigrants, tourists, visitors, and strangers who can access both the direction and voice narration provided by Google Maps in a language of choice to ease navigation. This automatic navigation service with advanced features such as street view, location of landmarks and significant locations as well as the calculation of best route, determination of traffic congestion, and prediction of the estimated time of arrival [1] to a predetermined route has made traveling arguably universally more interesting. The voice narration offered by Google Maps makes the navigation experience seamlessly pleasant or close to it. This navigation service and many such related assistive technologies however have one major drawback, they are primarily designed in English language with support for only a handful of languages thereby cutting off a significant number of users who do not have the command of the supported languages as the case is in Africa. Getting a translation of the service is dicey as well, as [2] submits that publicly available voice translation models (provided by Google) only support a range of 21 to 113 source languages. Of the 133 languages supported by Google, only 32 are available in voice or conversation mode. Of these 32 languages, only 4 are African languages, namely, Afrikaans, Amharic, Hausa, and Swahili.

Yoruba is one of the most popular languages in Africa. It is native to the southwestern region of Nigeria (the most populous African country), with more than thirty million speakers [3]. Yoruba dialects are also spoken in countries like Togo and Benin Republic. Outside Africa, Yoruba speakers are found in countries such as Trinidad and Tobago, Cuba, and Brazil. Advances in the field of natural language processing, particularly speech translation can be leveraged to widen the reach of assistive technologies such as navigation services narration in a language such as Yoruba.

Translation, which plays an indispensable role in globalization, is simply the art of making a sentence or speech in a source language available in another language (a target language) in a manner comprehensive enough for the intended audience. Speech Translation thus seeks to automate the process of translating spoken words from one language into another, which in effect bridges the language divide among the identified 7168 living languages globally [4]. The process of speech translation involves taking information about conversational speech in one language and using it to understand speech phrases in another language. There are three distinct methods involved in automatic speech translation technology: the ability to detect (speech acknowledgment), the ability to interpret (language interpretation), and the ability to integrate (speech union) one's speech into another language. Today, Speech Translation often involves an automatic recognition of a given speech, the automatic translation of the speech to a text form which is then automatically read out (spoken) by a machine.

This paper ultimately aims for a Yoruba-speaking Google Map system, by seeking an effective translation of the Google Map direction narration, which is primarily generated in English, to the Yoruba Language.

II. Review of Related Works

The ability to communicate and pass on information is a primary feature of the human race and arguably the pivot on which human existence stands. Over time, there has been a serious effort to enhance seamless verbal communication among humans of different races and origins. Speech-to-Speech Translation (STST) models greatly facilitate such efforts. STST originated from a series of interconnected (cascade) systems comprising Automatic Speech Recognition (ASR), Machine Translation (MT), and text-to-speech (TTS) synthesis [5], [6].

There have been significant improvements in research on direct STST which eliminates the dependency on MT-generated texts and the attending constraints. The applications of direct STST are however still limited hence the exploration of a cascaded system for this translation of English narration to a Yoruba narration.

A. ASR

The utilization of ASR has significantly boosted the development of translation. ASR is the method by which a

continuous stream of speech is transformed into a text sequence composed of individual words. A specialized form of digital signal processing that involves the application of fields such as statistics, and linguistics, ASR is the process of converting speech or audio waves into machine-readable texts by analyzing and processing speech signals using various techniques [7]. An ASR system therefore can detect provided verbal or speech input accurately, recognize the uttered words with precision, and subsequently utilize the recognized words as input for another machine to execute a certain operation [8].

ASR reinforces the human need for verbal communication which has been the most innate, effective, and best method of interaction among individuals until the advent of computers which necessitates providing input and commands nonverbally into diverse technological devices. With ASR, speech input into computing devices has become possible.

ASR has been in development for years but has made notable significance over the past decade. [9], [10] submit that conventional ASR models were built on the Gaussian mixture model-hidden Markov model, a fundamentally generative model through which speech signals are generated using Bayesian inference to determine the uttered words. There were several modifications to this model over time with varying degrees of success. In recent times, deep learning techniques such as Deep Neural Network Hidden Markov Model and End-to-End (E2E) models have become popular, showing remarkable achievements in the field of ASR.

Today, ASR, particularly for high-resourced languages, has made substantial advancements in performance and is leveraged in various applications and fields such as voice assistants, robotics, education, and search engines, to name a few. The same, however, cannot be said of low-resourced languages.

Applying natural language processing to low-resource languages is hindered by numerous constraints. The first, and perhaps the most significant, is the divergent morphological setup of different languages. The predominant language used as the foundation for most natural language processing techniques is English. English is regarded as having a simple morphology

[11] in stark contrast to many other languages that possess complex morphological structures with extensively inflected and polysemic words [12], [13]. Therefore, the methods that have been used successfully in English cannot adequately handle languages that have rich morphology structure. Numerous studies, such as those by [14], [15], [3] and others, have focused on ASR for tonal languages, and African languages.

There has been an increasing interest in ASR for low resource languages to make NLP capabilities available for these less-represented languages [16], and E2E models have shown remarkable results, although the results are still far from perfect. E2E models require extensive resources and substantial amounts of labeled speech data, (especially transcribed speech) which most spoken languages lack, to achieve excellent performance.

The transfer learning paradigm which involves the training of a model on labeled speech data from one or more high resource languages, and subsequently adjusting or refining the model using speech data from a low-resource language has been greatly explored [17]. The various nuances and complexity in the grammar formation and vocabularies of these divergent languages however affect the output of these E2E low-resource ASR models such that [18] submits that the most advanced ASR models are constrained when it comes to low-resourced languages. [19] identified limited availability of speech and text data, absence of standardization with variations in pronunciation as well as the unique properties each language possesses as shown in their linguistic and phonetic composition as the three primary challenges ASR for low-resourced languages face. To produce accurate transcripts from speech patterns identified in a specific input language speech, ASR models require a substantial amount of training data. Addressing these challenges has been the focus of much research lately.

B. Machine Translation

In this cascaded model of STST, speech recognition is often followed by the translation of the transcript to a target language which is then utilized in subsequent speech generation. This translation process, referred to as “Machine Translation” (MT) is a completely automated process that transforms a source language into the desired target language, a significant endeavor that seeks to utilize computers in the translation of sentences in natural language.

[20] stratified modes of machine translation into four: Rule based, corpus-based, Hybrid, and knowledge-based MT models. The initial methodology for machine translation mainly depends on manually created translation rules and

linguistic expertise [21]. Due to the intrinsic complexity of natural languages, it is challenging to account for all linguistic inconsistencies using manual translation methods [22], this is in addition to the exorbitant cost of maintenance and updating [23], all of which underscores the limitations of rule-based MT models. Knowledge-based MTs are made up of a vast repository of parallel texts and an inference engine. The knowledge-based MT models are however fraught with the complexity of representing knowledge and establishing its level of detail. The hybrid approach to machine translation combines two or more MT techniques. While this approach is adjudged to produce considerably optimal results, it is costly to implement [20].

The availability of extensive parallel corpora led to a rise in interest in data-driven methods that extract linguistic information from the data from which Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) were birthed. SMT differs from rule-based machine translation in that it acquires hidden structures, such as word alignments or phrases, directly from parallel corpora. The efficacy of SMT systems is directly influenced by the number of parallel sentence pairs accessible for training [24]. The translation quality of SMT is reported in [22] as unsatisfactory due to its inability to represent long-distance connections between words. Today, NMT is the dominant approach in the field of machine translation. NMT, designed to significantly depend on the use of encoders and decoders, is a translation system that uses a completely automated neural network [25]. Instead of translating individual words separately, NMT achieves a higher level of accuracy by taking into account the surrounding context in which the words are utilized. NMT aims to construct and train a single, extensive neural network that can accurately translate a sentence [23].

Many NMT models are available that generate translations for sentences in high-resourced languages (such as English, German, and Spanish) inputted efficiently. Some of these models are developed by global technological giants including Google with Google Translate, Meta with the No Language Left Behind model, IBM with IBM Watson, etc. Noting that the many real-world uses of MT position it as the most prominent NLP application, [26] submits that when it comes to low-resource languages, MT systems continue to fall short. In corroboration, [27], [28] accede that the translation accuracy of MT models for low-resource languages which often have complex morphology remains inadequate. NMT models operate on the assumption that there is an ample amount of bilingual training data (between a high-resource language and a targeted low-resource language), however, this is rarely the case in real-world scenarios [29]. Therefore, the output of the models built remains far from being desirable. Indeed, the majority of widely used translation engines do not even support a large number of low-resource languages. Improvements in translation quality for native languages spoken by a large portion of the world's population would bring the impact of MT to bear.

C. Text to Speech Translation

Using written text as input, Text to Speech Translation also known as speech synthesis is the process of generating artificial accents that closely resemble human speech. The advent of deep learning has enabled the most sophisticated speech synthesis systems to produce speech that is exceptionally natural-sounding and straightforward to comprehend [30].

Deep Neural Network [31], [32] revolutionized TTS, producing remarkable results in human speech synthesis, a huge leap from the once popular statistical parametric speech synthesis models [33]. Deep learning however relies heavily on a substantial quantity of training data [34], [35] such that it was stated in [33], [36] that DNN is not a suitable technique for TTS in low-resource languages. In [37] however, techniques such as monolingual transfer learning, cross-lingual transfer learning, multi-speaker models, multilingual models, and data augmentation have been proposed as means of augmenting TTS for low-resource languages.

D. Limitations of Cascaded Speech to Speech Translation Approach

There are several drawbacks to cascaded STS systems, even though they can be constructed on top of the current concurrent components. The most notable of the downside of the cascaded STS translation is dependency on intermediary textual outputs which constitutes constraints for these cascaded models in facilitating efficient inference and unwritten languages. [38] noted that there is an increase in latency in cascaded STS models due to the pipeline of many models. This latency originates either from the time it takes for numerous models to compute or the delay that results from their processing not being in sync with one another.

To mitigate the constraints posed by cascaded STS systems, many research focuses have turned to direct methods for speech translation that do not require written texts. Two distinct models of STS translation were developed: the first, a Sequence-to-Sequence Translation (S2ST) of the source to the target language, and the translation of the source to the target language as a discrete unit. Google's Translatotron [39] and Translatotron2 [40] employ the end-to-end (S2ST) model [41] to produce target spectrograms via multitask learning. Another avenue of investigation is substituting the desired spectrograms in S2ST modeling with distinct units that are acquired through extensive analysis of unannotated speech [42]–[44]. Discrete units have been shown to demonstrate a superior ability to capture language content compared to spectrograms. Although there have been advancements in direct S2S translation, a major obstacle that remains is the scarcity of parallel speech data.

III. Proposed System

This section describes the setup of the proposed English-to-Yoruba narration translation system.

The STS Translation model proposed is made up of three key components: an ASR model to get the voice from the Google Maps direction narration, an MT model for translating the voice to the desired language (Yoruba), and a TTS model for producing the desired narration in the target language. In this setup, state-of-the-art models for each of the three stages of the cascaded speech-to-speech translation system were thoroughly explored to measure their suitability for the task before coalescing the different models together into a single system. A pictorial representation of the proposed model is given in Figure 1.

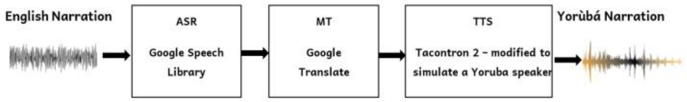


Fig. 1. The translation model

Several publicly available models for each of the three cascaded stages in this translation exercise as specified in the methodology (ASR, MT, and TTS) were explored. In addition to being publicly available, the models were chosen based on their popularity and widely reported performance. The models were then assessed based on the Word Error Rate (WER) performance.

WER is a widely used measure to evaluate the accuracy of an automatic speech recognition system. It calculates the combined number of words that are deleted, substituted, or inserted in given sentences by a machine translation model. WER is calculated as given in equation 1.

$$WER = \frac{S + D + I}{N} \tag{1}$$

The main challenge in measuring performance with WER is the discrepancy in length between the machine-recognized word sequence and the reference (primary) word sequence [45]. Calculated based on the Levenshtein distance [46], WER operates at the level of individual words rather than phonemes which does not offer any specific information regarding the specific types of translation error. [47] submits that when the WER is high, it becomes challenging to determine the level of usefulness a model will have for the end user. Specifically, when WER fails to account for the degree of semantic similarity or the evaluation of the mistake. The availability of other robust metrics for ASR model evaluation does not however invalidate or make WER evaluation trivial.

A. Source/Input

The input in this Yoruba-speaking Google Maps direction narration system is the narration (in the English language) provided by the Google Maps direction service recorded as an audio file.

B. ASR

Several open-sourced ASR models were considered for the task of generating the transcript (text input) from the audio source input. Five of these models were explored including WhisperAI, developed by Open AI, Facebook's FAIRSEQ S2T (S2T-LARGE-LIBRISPEECH-ASR model) [48], Microsoft's unispeech-sat-base-100h-libri-ft ASR model [49], Nvidia's Conformer-Transducer X-Large STT [50] and the Google Speech Library. These models possess the capability of audio capture, speech recognition, and transcript generation.

When evaluating these models, we took into consideration, the submission in [51] which highlights the challenges that impact the quality of speech recognition. These challenges range from interference from background noise, variations in accents, dialects, word length, and non-wordy utterances like breaths, coughs, or sneezes. We therefore evaluated the models in a serene environment as well as in a relatively noisy environment.

The output of the models as well as their performance based on their WER evaluation are presented in Tables I to IV.

TABLE I
ASR MODEL PERFORMANCE IN NOISY ENVIRONMENT

Reference Text	In two hundred meters turn left
ASR Model	Output
WhisperAI	*In 200 meters turn left.*
FAIRSEQ S2T (S2T-LARGE-LIBRISPEECH-ASR)	*and two hundred metres on the left wall*
Microsoft's unispeech-sat-base100h-libri-ft ASR	*AND GON ADMI DAZ DON ALETO U*
Nvidia's Conformer-Transducer XLarge STT	time-out error
Google Speech Library	*let in 200 m turn left*

Table 1 shows the output of each of the models in a relatively noisy environment. Only Nvidia's conformer-transducer STT model failed to produce an output in the given experimental setup, all other models responded to speech input and generated an output.

TABLE II
WER PERFORMANCE OF THE ASR MODELS IN NOISY ENVIRONMENT

Model performance					
ASR Model	WER	Substitutions	Insertions	Deletions	hit
WhisperAI	0.33	1	0	1	4
FAIRSEQ S2T (S2T-LARGELIBRISPEECH-ASR)	0.83	2	0	2	4
Microsoft's unispeech-satbase-100hlibri-ft ASR	1.17	6	1	0	0
Nvidia's ConformerTransducer X-Large STT	*not rated*	*not rated*	*not rated*	*not rated*	*not rated*
Google Speech Library	0.67	4	0	0	2

The WER of each model judging by the number of words substituted, inserted, and deleted is presented in Table II. From the results presented, WhisperAI model has the best WER score with one of the words in the reference text deleted, one of the words substituted, and none of the words substituted. There was a hit on four of the six words in the reference text. A similar performance was recorded by Facebook's Fairseq S2T model, obtaining a hit on four of the six reference texts. Google Speech library substituted four of the six reference texts while retaining the meaning of the words. Microsoft's unispeech-sat-base -100h-libri-ft ASR substituted all the words in the reference text without retaining the meaning of any of the words. The pictorial representation of the performance of these models is shown in Figure 2.

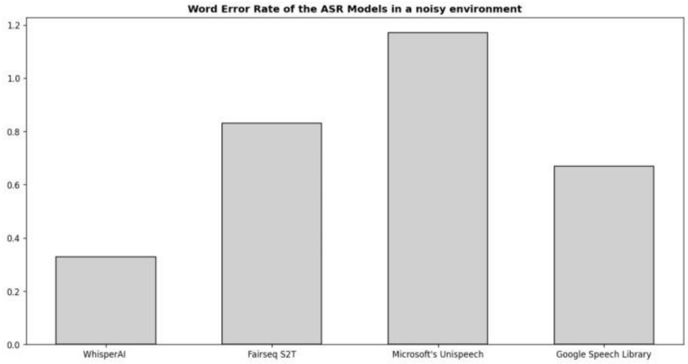


Fig. 2. The Word Error Rate of the models in a noisy environment

TABLE III
ASR MODEL PERFORMANCE IN SERENE ENVIRONMENT

Reference Text	In two hundred meters turn left
ASR Model	Output
WhisperAI	*In 200 meters turn left.*
FAIRSEQ S2T (S2T-LARGE-LIBRISPEECH-ASR)	*in two hundred metres tom leaped*
Microsoft's unispeech-sat-base100h-libri-ft ASR	*IN TWO HUNDRED MITALS JON LEPTD*
Nvidia's Conformer-Transducer XLarge STT	time-out error
Google Speech Library	*in 200 m turn left*

In Table III, the performance of the models was examined again in a serene environment void of background noise. Again, the WhisperAI has the best WER result with four hits of the six reference texts. Facebook's Fairseq S2T had the second-best WER result with four hits out of the six reference words. It, however, substituted two words "turn left" for "tom leaped", which shows that the information the sentence is passing across has been lost and underscored the need for more evaluation of ASR models other than WER. While Google Speech Library also substituted two words, "two hundred meters" to "200 m", it still retained the meaning of the sentence. The Microsoft ASR model examined had three words substituted and got three-word hits; it however does not retain the information the reference sentence passes across. Of the five models (apart from Nvidia's model which returned an error), only whisperAI and Google Speech Library ASR models were able to retain the meaning or interpretation of the reference sentence.

TABLE IV
WER PERFORMANCE OF THE ASR MODELS IN A SERENE ENVIRONMENT

Model performance					
ASR Model	WER	Substitutions	Insertions	Deletions	hit
WhisperAI	0.33	1	0	1	4
FAIRSEQ S2T (S2T-LARGELIBRISPEECH-ASR)	0.5	2	0	0	4
Microsoft's unispeech-satbase-100hlibri-ft ASR	0.67	3	0	0	3
Nvidia's Conformer-Transducer X-Large STT	*not rated*	*not rated*	*not rated*	*not rated*	*not rated*
Google Speech Library	0.67	2	0	1	3

The pictorial representation of the word error rate of the models assessed in a serene environment is presented in Figure 3.

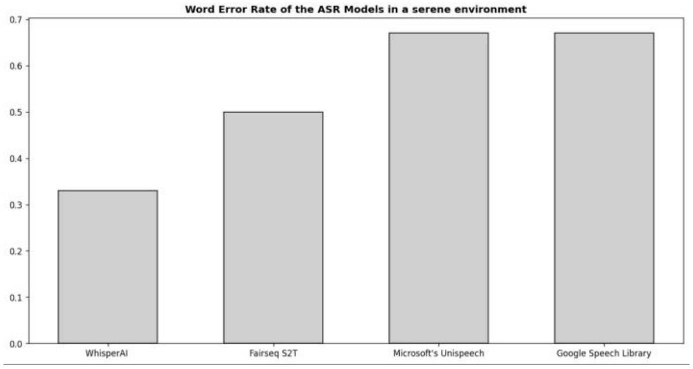


Fig. 3. The Word Error Rate of the models in a serene environment

C. Machine Translation

The next step in this cascaded speech-to-speech translation model is the translation of the text generated from the audio input in the ASR stage. Four models were explored in this stage.

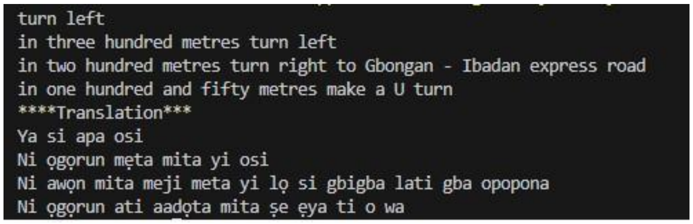
1) DeepL Translate: The DeepL GmbH team created the widely used DeepL machine translation service. The ability to translate text between numerous languages with ease and speed is provided by this free online application. The model uses deep learning techniques, especially transformer models which are trained on massive volumes of multilingual text data, DeepL can grasp linguistic subtleties and generate remarkably accurate and coherent translations. Unfortunately, DeepL does not support translation in Yoruba Language.

2) LibreTranslate: LibreTranslate is an open-sourced MT technology that offers translation services across several languages. The API offers a straightforward and user-friendly interface that can be seamlessly incorporated into different applications. LibreTranslate however does not support translation in the Yoruba language.

3) NLLB: No Language Left Behind (NLLB), research championed by Meta to narrow the disparity between well-resourced and under-resourced languages in the field of machine translation, provides translation services for an impressive range of 200 languages, including some African languages that are frequently overlooked by other models. This feature renders it an invaluable instrument for facilitating worldwide communication and enhancing the availability of information. It employs innovative methods to train with a small amount of data and enhance the quality of translation for various languages. The NLLB model is programmed to recognize the target language and acquire proper syntax and grammar. Data collected by social media platforms like Facebook is utilized to train this model, leading to the creation of many language training datasets [5]. Research indicates that NLLB provides high-quality translations, with a focus on achieving precision and producing output that sounds natural. It does not however support translation in the Yoruba language.

4) Google Translate: Google Translate, the company's language translation product, debuted in 2006 and initially supported only two languages using their statistical machine translation engine [52]. In order to steadily improve the quality of its translations, Google Translate now uses machine learning and neural network technologies. It currently does hundreds of millions of translations every day, earning it a reputation as the most prominent provider of online language translation services. The Google Translate API is a Python-based library that dynamically converts text provided in a language to a desired language through the use of a neural machine translation that has been pre-trained. Google Translate translates text according to the frequency of occurrence of word pairs between two languages. The absence of context-aware translation may result in grammatical errors or the loss of meaning in the translated text. More than 100 different languages are currently supported by Google's extensive translation services. It facilitates the translation of the Yoruba language into other languages and vice versa.

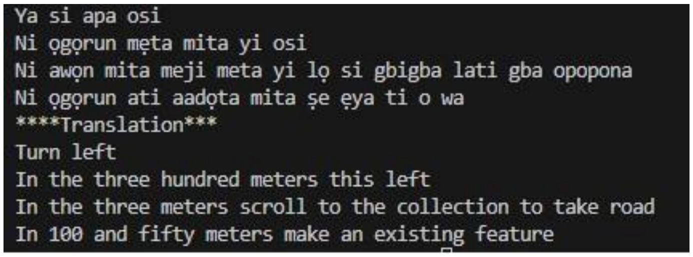
Only Google Translate supports translation into the Yoruba language of the four state-of-the-art neural machine translation models explored. The output of the transcript generated in the ASR part of this task, which is an English text document is therefore translated to Yoruba language using Google Translate. While the model was able to translate some of the texts correctly, a significant part of the translations was wrongly interpreted, especially locations in Yoruba and Yoruba numerals. An example of this is presented in Figure 4 and Figure 5 followed by the evaluation of the translations using the word error rate metric.



```
turn left
in three hundred metres turn left
in two hundred metres turn right to Gbongan - Ibadan express road
in one hundred and fifty metres make a U turn
****Translation***
Ya si apa osi
Ni ọgọrun mẹta mita yi osi
Ni awọn mita meji meta yi lọ si gbigba lati gba opopona
Ni ọgọrun ati aadọta mita ṣe ẹya ti o wa
```

Fig. 4. The Google translation of the ASR-generated text from English to Yoruba

Figure 4 shows the Google translation of the ASR-generated text from English to Yoruba. From Figure 4, using human evaluation of the MT-translation, three of the four sentences were not translated correctly. The total number of words in the reference text is 29, and the errors recorded from the substitutions, deletions, and wrongful insertions recorded by the model are 12. The WER evaluation of the translation is 41%. The Yoruba text obtained from this translation was then re-translated into English, using Google Translate. The result is presented in Figure 5.



```
Ya si apa osi
Ni ọgọrun mẹta mita yi osi
Ni awọn mita meji meta yi lọ si gbigba lati gba opopona
Ni ọgọrun ati aadọta mita ṣe ẹya ti o wa
****Translation***
Turn left
In the three hundred meters this left
In the three meters scroll to the collection to take road
In 100 and fifty meters make an existing feature
```

Fig. 5. The Google translation of the MT-generated Yoruba text to English

TABLE V
EVALUATION OF THE MACHINE TRANSLATION YORUBA TEXT

Google Transla:e Model Evaluation performance	
Reference Text	Output
turn left	*turn left*
in three hundred meters turn left	*in three hundred meters this left*
in two hundred metres turn right to Gbongan - Ibadan express road	*in the three meters scroll to the collection to take road*
in one hundred and fifty metres make a U-turn	*in 100 and fifty meters make an existing feature*

Table V contains four sentences that make up the reference translation task containing 29 words. In the translation, there were 12 words substituted, 2 words deleted, no wrong word inserted, and 17 words gotten correctly. The WER for this translation task stands at 48%.

TABLE VI
WER EVALUATION OF TABLE V

WER of each sentence					
Sentence	Word Length	Substitutions	Insertions	Deletions	hit
1	2	0	0	0	2
2	6	1	0	0	5
3	11	7	0	1	3
4	10	4	0	1	5

In the end, the transcript of the audio input generated through ASR is:

> ”turn left, in three hundred meters turn left, in two hundred meters turn right to Gbongan - Ibadan express road, in one hundred and fifty meters make a U-turn”

The Yoruba translation of the task is:

> ”Ya si apa osi, ni ogorun meta mita yi osi, ni awon mita meji meta yi lo si gbigba lati gba opopona, ni ogorun ati aadota mita ṣe eya ti o was”

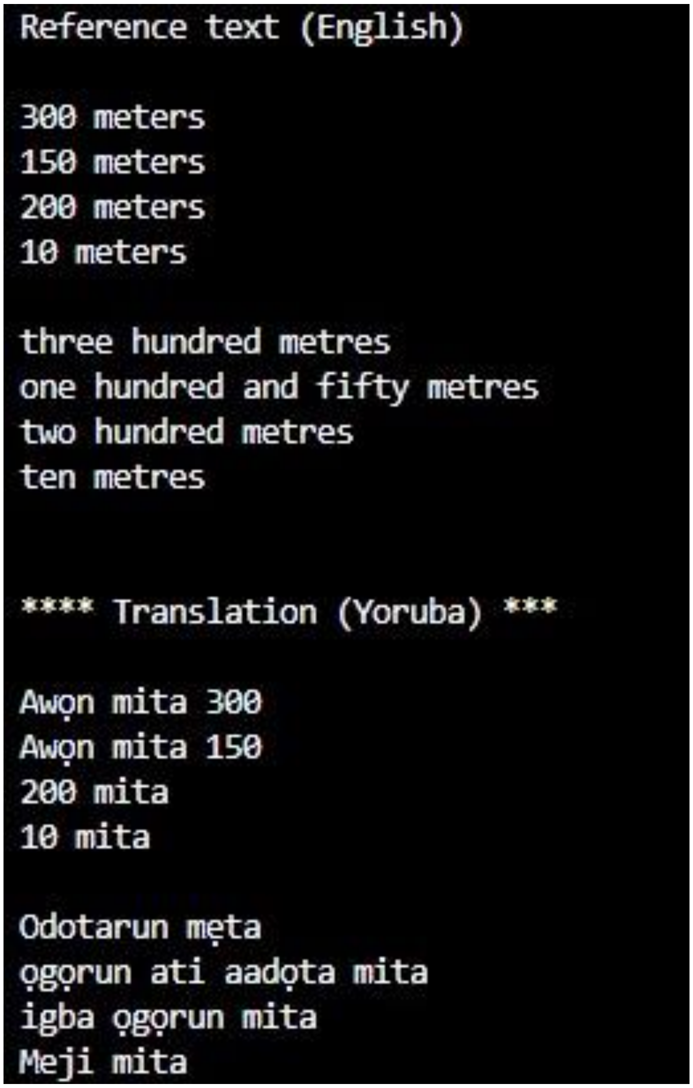
This Yoruba translation, when translated to English using Google Translate is:

> ”turn left, in three hundred meters this left, in the three meters scroll to the collection to take road, in 100 and fifty meters make an existing feature”.

It is clear from the results presented that the information contained in the input text has been lost in translation. It is therefore safe to submit that this speech-to-speech translation model so far cannot provide a Yoruba translation reliable enough to build a Yoruba-speaking Google Map navigation narration system upon without some modifications or improvement.

It can be inferred that the bulk of the error this translation model recorded revolves around Yoruba numerals and locations which form the pivot of this translation exercise. We therefore probed a little further, limiting the reference text to be translated using GoogleTrans to numerals only. The output is presented in Figure 6.

The reference text as presented in Figure 6 contains numbers written in both figures and words in a bid to ascertain the performance of the model in the translation of the various forms in which numbers can be shown. We compared the translation output with the expected (human) translation and presented the result in Table VII. There was a wrongful insertion in two of the four reference texts presented as figures. The model was marked with wrong substitutions when translating numerals presented in words. For instance, ten metres, which is expected to be *ibuso mewa*, was translated as *meji meta*. While the word *meter* has been substituted as *mita* in the translation instead of *ibuso*, it is still acceptable under the premise of lexical borrowing [53], [54]. However, the substitution that occurs in the translation of Yoruba numbers such as *three hundred* to *odotarun meta* instead of *orundin nirinho* completely contravenes either of the Yoruba numerals formation - the cardinal (*asoye or onkaye*) or the ordinal (*asopo or onkapo*) system [55], [56]. We argue that the word *odotarun* does not exist in the Yoruba language. Interestingly, when *odotarun meta* was translated to English using Google Translate, it returned three hundred and fifty. There is therefore something fundamentally wrong either with the dataset or the techniques on which the model has been configured to handle Yoruba numerals.



```
Reference text (English)

300 meters
150 meters
200 meters
10 meters

three hundred metres
one hundred and fifty metres
two hundred metres
ten metres


**** Translation (Yoruba) ***

Awọn mita 300
Awọn mita 150
200 mita
10 mita

Odotarun mẹta
ọgọrun ati aadọta mita
igba ọgọrun mita
Meji mita
```

Fig. 6. The translation of some Numerals from English to Yoruba using Google Translate

TABLE VII

Evaluation of the machine translation of Yoruba numerals

Human Translation of Yoruba numerals vs translation by Google Translate		
Reference Text	Expected (Human) Translation	Expected (Human) Translation
300 meters	*ibuso orundin-nirinho*	*awon mita 300*
150 meters	*ibuso aadojo*	*awon mita 150*
200 meters	*ibuso igba*	*200 mita*
10 meters	*ibuso mewa*	*10 mita*
three hundred metres	*ibuso orundin-nirinho*	*odotarun meta*
one hundred and fifty metres	*ibuso aadojo*	*ogorun ati aadota mita*
two hundred metres	*ibuso igba*	*igba ogorun mita*
ten meters	*ibuso mewa*	*meji mita*

D. TTS

Five Text-To-Speech models were explored in this next phase of the translation exercise. The input to this TTS is the Yoruba machine-translated text from the MT section. The models are thus presented.

1) Seamless TTS: Developed by Meta, SeamlessM4T v2 is a comprehensive TTS model capable of processing both speech and text singly. It supports 100 languages for speech input and 96 languages for text input. The system can generate written output in 96 languages and spoken output in 36 languages [2]. Unfortunately, Yoruba is not one of the languages supported for the spoken output.

2) MMS TTS: The Massively Multilingual Speech TTS model, an endeavor by Facebook at significantly increasing speech technology to more than 1000 languages, providing TTS models for 1107 languages [57], including Yoruba. The mms-tts-yoruba model was able to convert the input texts to Yoruba words. However, the model still needs fine-tuning. *3) SpeechT5 TTS:* The Microsoft speech synthesis model [58] does not support TTS in Yoruba.

4) Coqui TTS: Coqui promises several noteworthy features, including voice cloning, and multilingual speech generation capabilities [?]. It however supports only 16 languages, none of which is an African language.

5) Your TTS: The YourTTS model [59], a multi-speaker speech synthesis model does not also support Yoruba language.

Only one of the five state-of-the-art TTS models that were investigated to be utilized for the third and final stage of this English narration to Yoruba narration translation gave a perceptible outcome. This model was Facebook's Massively Multilingual Speech TTS.

IV. Conclusion

We have examined the various components that make up a cascaded speech-to-speech translation model aimed at facilitating the translation of the voice narration provided by Google Maps direction service from English to narration in the Yoruba language.

The first of the three components is the ASR. Four of the five models experimented with at this stage were able to conveniently recognize the English voice narration and produce a corresponding text with varying degrees of accuracy. The WhisperAI model was adjudged the best of these models based on a WER evaluation score of 33%. Although a WER of 33% is considerably high (i.e. the proportion of words that the model got wrong when compared to the words that were actually spoken), this performance is deemed acceptable because contextually, the meaning of the words spoken remained intact at the end of the translation.

The second component is the MT: i.e. the automatic translation of the English text generated from the ASR into Yoruba. We experimented with four state-of-the-art models of which only one, Google Translate, can translate English texts to Yoruba and vice versa. The translation was however fraught with significant errors especially when it came to Yoruba numerals and locations. We submit that the output (the Yoruba text generated during translation) at this stage can not be used as the input for the TTS model because the information from the reference text has become mutilated and lost in the previous translation stages.

The last component of the model is the TTS: i.e. the automatic production of the Yoruba speech from the translated texts. We experimented with five state-of-the-art TTS models for this exercise and only one model, Meta's Massively Multilingual Speech TTS model supports TTS in the Yoruba language. The model however needs some fine-tuning to get the pronunciation of the words to a largely acceptable level.

V. FUTURE WORK

The findings from this work reveal that while there is considerable effort in the automatic speech-to-speech translation especially for low-resource languages, the task of a speech-to-speech translation in the Yoruba language lags conspicuously on two fronts: an efficient machine translation model and a formidable text-to-speech model for speech synthesis.

Here, we summarize the direction in which we may turn our future efforts.

1) Development of a high-performing Machine Translation model for the Yoruba language: While we acknowledge the great effort in state-of-the-art models for translation to and from Yoruba, we believe that there is still room for improvement. The best of the models we examined had a word error rate of 30% in short and basic sentences. The word error rate would only be significantly worse while using such models in translating long sentences. Because we have a first-hand understanding of the nuances of the language and understand its linguistic and semantic structure, we, therefore, intend to develop a Yoruba machine translation model that will have better performance and will be publicly available.

2) Curation of more publicly available audio dataset for the Yoruba language: Advanced machine translation models today are primarily transformer-based models which often require a substantial amount of training data. In the course of this study, we found three publicly available Yoruba audio datasets on the internet. The first, OpenSLR86 a 4-hour-long transcribed dataset provided by OpenSLR and Google; the second, the Lagos-NWU, a 2 hours 45-minute long audio file transcribed provided by North-West University; and the Bibeli Mimo (NIV) a 93:38:15 audio file transcribed, provided by Biblica Open Bible. Apart from the Bibeli Mimo dataset which is fairly long, the other datasets combined provide an approximate 7 hours dataset which is significantly small in the context of this assignment. The Bibeli Mimo dataset, being a religious text however, will be highly constrained and limited largely to religious registers. All these underscore the need for more audio datasets for training Yoruba TTS systems.

3) Development of formidable Text to Speech model for the Yoruba Language: We believe that for the TTS task, the Tacotron2 model can be adopted for generating the Yoruba narration of the Google Map Direction translation from English to Yoruba. The model was primarily trained on a sample of US accents, which will not suit this paper's target language, Yoruba, which cannot contain any foreign accent. The Tacotron2 is capable of being trained on a single language, however, it is not suitable for use in cross-lingual text-to-speech systems. Therefore, it is necessary to train a model in which the acoustic and language properties are trained independently. Thus, the need for a where the acoustic and linguistic features are separately trained. To produce a proper Yoruba accent that has both acoustic and linguistic features intact, a model can be trained adopting the architecture in [5] using any of the publicly available Yoruba datasets or a newly curated one.

References

[1] H. Mehta, P. Kanani, and P. Lande, "Google maps," *International Journal of Computer Applications*, vol. 178, no. 8, pp. 41–46, 2019.

[2] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[3] O. Adetunmbi, O. O. Obe, and J. Iyanda, "Development of standard yorub` a speech-to-text system using htk,"´ *International Journal of Speech Technology*, vol. 19, pp. 929–944, 2016.

[4] G. F. S. Eberhard, David M. and C. D. F. (eds.), "Ethnologue: Languages of the world. twenty-sixth edition. dallas, texas: Sil international.," 2023. Accessed on January 26, 2024.

[5] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswani, C. Wang, J. Pino, B. Sagot, and H. Schwenk, "Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations," *arXiv preprint arXiv:2211.04508*, 2022.

[6] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speechto-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 958–965, IEEE, 2021.

[7] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, "Arabic automatic speech recognition: a systematic literature review," *Applied Sciences*, vol. 12, no. 17, p. 8898, 2022.

[8] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[9] D. Yu and L. Deng, *Automatic speech recognition*, vol. 1. Springer, 2016.

[10] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[11] A. Chakrabarty and U. Garain, "Benlem (a bengali lemmatizer) and its role in wsd," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 15, no. 3, pp. 1–18, 2016.

[12] O. Dereza, "Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of early irish," in *Proceedings of Third Workshop Computational linguistics and language science*, vol. 4, pp. 113–124, 2019.

[13] A. Gogoi and N. Baruah, "A lemmatizer for low-resource languages: Wsd and its role in the assamese language," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 4, pp. 1– 22, 2022.

[14] F. H. Khosoa, S. Z. Nasira, and D. N. Hakroc, "Challenges of accent and vowels for sindhi speech recognition system," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, p. 916–921, Apr 2021.

[15] J. Kaur, A. Singh, and V. Kadyan, "Automatic speech recognition system for tonal languages: State-of-the-art survey," *Archives of Computational Methods in Engineering*, vol. 28, pp. 1039–1068, 2021.

[16] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7909–7913, IEEE, 2020.

[17] S. Khare, A. R. Mittal, A. Diwan, S. Sarawagi, P. Jyothi, and S. Bharadwaj, "Low resource asr: The surprising effectiveness of high resource transliteration.," in *Interspeech*, pp. 1529–1533, 2021.

[18] M. Baas and H. Kamper, "Voice conversion can improve asr in very low-resource settings," *arXiv preprint arXiv:2111.02674*, 2021.

[19] W. Du, Y. Maimaitiyiming, M. Nijat, L. Li, A. Hamdulla, and D. Wang, "Automatic speech recognition for uyghur, kazakh, and kyrgyz: an overview," *Applied Sciences*, vol. 13, no. 1, p. 326, 2022.

[20] S. A. B. Andrabi *et al.*, "A review of machine translation for south asian low resource languages," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1134–1147, 2021.

[21] S. K. Pulipaka, C. K. Kasaraneni, V. N. S. Vemulapalli, and S. S. M. Kosaraju, "Machine translation of english videos to indian regional languages using open innovation," in *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pp. 1–7, IEEE, 2019.

[22] Z. Tan, S. Wang, Z. Yang, G. Chen, X. Huang, M. Sun, and Y. Liu, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, 2020.

[23] I. Rivera-Trigueros, "Machine translation systems and quality assessment: a systematic review," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593–619, 2022.

[24] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.

[25] K. Dedes, A. B. P. Utama, A. P. Wibawa, A. N. Afandi, A. N. Handayani, and L. Hernandez, "Neural machine translation of spanish-english food recipes using lstm," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 2, pp. 290–297, 2022.

[26] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzman, and A. Fan, "The flores-101 evalu-´ ation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022.

[27] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation," *Tsinghua Science and Technology*, vol. 27, no. 1, pp. 150–163, 2021.

[28] L. S. Meetei, S. M. Singh, A. Singh, R. Das, T. D. Singh, and S. Bandyopadhyay, "Hindi to english multimodal machine translation on news dataset in low resource setting," *Procedia Computer Science*, vol. 218, pp. 2102–2109, 2023.

[29] J. Zhang and C. Zong, "Neural machine translation: Challenges, progress and future," *Science China Technological Sciences*, vol. 63, no. 10, pp. 2028–2050, 2020.

[30] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, 2022.

[31] K. Azizah and W. Jatmiko, "Transfer learning, style control, and speaker reconstruction loss for zero-shot multilingual multi-speaker texttospeech on low-resource languages," *IEEE Access*, vol. 10, pp. 5895–5911, 2022.

[32] A. Elneima and M. Binkowski, "Adversarial text-to-speech for low-´ resource languages," in *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp. 76–84, 2022.

[33] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on lowresource languages," *IEEE Access*, vol. 8, pp. 179798–179812, 2020.

[34] A. R. Gladston and K. V. Pradeep, "Exploring solutions for text-tospeech synthesis of low-resource languages," in *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, pp. 168–172, 2023.

[35] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi, and H. Saruwatari, "Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining," *arXiv preprint arXiv:2301.12596*, 2023.

[36] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6940–6944, IEEE, 2019.

[37] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 42, 2021.

[38] X. Ma, H. Gong, D. Liu, A. Lee, Y. Tang, P.-J. Chen, W.-N. Hsu, P. Koehn, and J. Pino, "Direct simultaneous speech-to-speech translation with variational monotonic multihead attention," *arXiv preprint arXiv:2110.08250*, 2021.

[39] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[40] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *International Conference on Machine Learning*, pp. 10120– 10134, PMLR, 2022.

[41] S. Zhang and Y. Feng, "End-to-end simultaneous speech translation with differentiable segmentation," *arXiv preprint arXiv:2305.16093*, 2023.

[42] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[43] A. Tjandra, S. Sakti, and S. Nakamura, "Speech-to-speech translation between untranscribed unknown languages," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 593–600, IEEE, 2019.

[44] A. Lee, P.-J. Chen, H. Schwenk, J. Gu, and W.-N. Hsu, "Textless speechto-speech translation on real data," June 15 2023. US Patent App. 17/889,116.

[45] A. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.," 01 2004.

[46] A. Trabelsi, S. Warichet, Y. Aajaoun, and S. Soussilane, "Evaluation of the efficiency of state-of-the-art speech recognition engines," *Procedia Computer Science*, vol. 207, pp. 2242–2252, 2022.

[47] J. Tobin, Q. Li, S. Venugopalan, K. Seaver, R. Cave, and K. Tomanek, "Assessing asr model quality on disordered speech using bertscore," *arXiv preprint arXiv:2209.10591*, 2022.

[48] C. Wang, W.-N. Hsu, Y. Adi, A. Polyak, A. Lee, P.-J. Chen, J. Gu, and J. Pino, "fairseq sˆ2: A scalable and integrable speech synthesis toolkit," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online and Punta Cana, Dominican Republic), pp. 143–152, Association for Computational Linguistics, Nov. 2021.

[49] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6152–6156, IEEE, 2022.

[50] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, *et al.*, "Nemo: a toolkit for building ai applications using neural modules (2019)," *arXiv preprint arXiv:1909.09577*, 1909.

[51] M. Stenman, "Automatic speech recognition an evaluation of google speech," 2015.

[52] O. Akinwale, A. Adetunmbi, O. Obe, and A. Adesuyi, "Web-based english to yoruba machine translation," *International Journal of Language and Linguistics*, vol. 3, no. 3, pp. 154–159, 2015.

[53] M. E. Oyinloye *et al.*, "Monophthongization in the adaptation of selected english loanwords in yoruba: A constraint-based analysis," *Journal of Universal Language*, vol. 21, no. 1, pp. 29–67, 2020.

[54] W. Osisanwo, O. Aina, and E. T. Bolaji, "Morphophonemics of yorub` a´ borrowed nouns in standard british english," *JOURNAL OF LINGUISTICS, LANGUAGE AND IGBO STUDIES (JoLLIS)*, vol. 3, no. 1, 2022.

[55] A. Adetomiwa, "YorUb` A numeral system in 21st century: Challenges´ and prospects," 10 2023.

[56] O. Babarinde, "Linguistic analysis of the structure of yoruba numerals," *Language Matters*, vol. 45, no. 1, pp. 127–147, 2014.

[57] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *arXiv*, 2023.

[58] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unifiedmodal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738, May 2022.

[59] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Golge, and¨ M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*, pp. 2709–2720, PMLR, 2022.