



# Extending the Fongbe to French Speech Translation Corpus: resources, models and benchmark

*D. Fortuné Kponou<sup>1</sup>, Salima Mdhaffar<sup>2</sup>, Fréjus A. A. Laleye<sup>3</sup>, Eugène C. Ezin<sup>1</sup>, Yannick Estève<sup>2</sup>*

<sup>1</sup>Institut de Mathématiques et de Sciences Physiques, Dangbo, Bénin

<sup>2</sup>LIA, Avignon, France

<sup>3</sup>OPSCIDIA, Paris, France

{fortune.kponou, eugene.ezin}@imsp-uac.org, {salima.mdhaffar,  
yannick.esteve}@univ-avignon.fr, frejus.laleye@opscidia.com

## Abstract

This paper introduces FFSTC 2, an expanded version of the existing Fongbe-to-French speech translation corpus, addressing the critical need for resources in African dialects for speech recognition and translation tasks. We extended the dataset by adding 36 hours of transcribed audio, bringing the total to 61 hours, thereby enhancing its utility for both automatic speech recognition (ASR) and speech translation (ST) in Fongbe, a low-resource language. Using corpus, we develop both cascade and end-to-end speech translation systems. Our models employ AfriHuBERT and HuBERT147 as encoders, and NLLB and mBART as decoders. We introduce a diacritic-substitution technique for both ASR and Machine Translation (MT) which, yields a BLEU score of 37.23 compared to 39.60 for the fully diacritized configuration. Among the evaluated end-to-end architectures, AfriHuBERT-NLLB with data augmentation attains the highest BLEU score of 26.32.

**Index Terms:** speech translation, speech recognition, African language, low resource language

## 1. Introduction

The creation of high-quality audio datasets for Natural Language Processing (NLP) tasks remains a significant challenge. Current efforts to develop speech datasets have predominantly focused on widely spoken languages such as English, French, and Spanish, leaving dialectal and minority languages largely under-represented. As a result, the vast majority of the world's 7,000 languages remain underserved, with only a few dozen language directions covered in existing speech translation corpora [1]. This lack of inclusion presents a critical problem, as it perpetuates language barriers and limits the accessibility of NLP technologies for speakers of less-represented languages. In recent years, there has been increasing attention on low-resource languages, particularly those spoken in regions such as India and Africa. Despite being spoken by millions of people, many of these languages remain severely under-represented in terms of linguistic resources. For instance, Africa is home to over 2,000 languages [2], yet, as highlighted by [3], only around 40 of these languages have been integrated into modern language technologies. This stark disparity underscores the significant under-representation of African languages in contemporary NLP research and applications.

Initiatives such as the Mozilla Common Voice project<sup>1</sup> have sought to address this gap by providing a platform for collecting speech data for certain African languages. However, the impact of such efforts remains limited. This limitation stems from the platform's design, which relies on collecting data through

the reading of written texts. This approach is less effective for many African languages, which are primarily oral and have limited written resources. Historically, oral traditions [4] have been widespread across the continent, which has hindered the development of writing systems and further complicated efforts to create comprehensive linguistic datasets.

While resources Fongbe are limited, some datasets do exist in the state of the art. For Fongbe, we have the ALFFA [5] dataset for speech recognition, which includes 6 hours of audio along with transcriptions, and the FFSTC dataset [6], which contains 31 hours of Fongbe speech paired with French translations. Also, corpora like GigaST are pseudo-labeled, distinguishing them from fully human-annotated datasets such as [7, 8] contained in ALFFA, which contains a few hours of data for languages like Amharic, Hausa, Swahili, Wolof.

In this paper, we present a significant extension of the existing Fongbe-to-French speech translation corpus [6]. This extension not only adds new parallel data from spoken Fongbe to written French but also enables the training of an automatic speech recognition (ASR) model for Fongbe by providing speech recordings with their corresponding transcriptions. This dataset will provide an opportunity for research community to test all new SSL models designed for African languages [9, 10], especially since no existing SSL model currently includes Fongbe in its training data. The data described here will be used to organize a translation task in IWSLT 2025<sup>2</sup>. This paper provides experiments and evaluation for Automatic Speech Recognition (ASR) in Fongbe and Speech Translation (ST) from Fongbe to French. These experiments aim to assess the effectiveness of leveraging pre-trained models for low-resource language processing, with a focus on a tonal African language spoken by 4 million people.

## 2. Related work and motivations

Unlabeled data is far more abundant than labeled data. Self-supervised learning (SSL) methods have emerged as a powerful approach to leverage such unlabeled data in machine learning, enabling the creation of pre-trained models. A first notable example in the domain of speech is the XLSR-53 model [11], which is pre-trained on 53 languages data. Studies [12, 13, 14] have extensively explored the use of pre-trained speech encoders and text decoders to enhance system performance for the speech translation task. These techniques, collectively referred to as transfer learning, have demonstrated significant effectiveness in improving performance, particularly in low-resource settings. By transferring knowledge from high-resource languages to low-resource ones, transfer learning provides a robust initialization for both encoder and decoder components,

<sup>1</sup><https://commonvoice.mozilla.org>

<sup>2</sup>International Workshop on Spoken Language Translation

thereby significantly contributing to improved translation accuracy. In addition to transfer learning, other approaches have been adopted to enhance performance in low-resource speech translation, such as synthesizing parallel data [15]. However, our work specifically focuses on the transfer learning paradigm, leveraging its proven capabilities to address the challenges of low-resource language processing.

Despite the challenges associated with creating speech corpora for African languages, there has been a notable shift towards the inclusion of these languages in pre-trained models. This trend is evident in the progressive integration of African languages into state-of-the-art models, such as HUBERT147 [10], which supports 16 African languages, and its variant AfriHuBERT [9], which extends coverage to 39 languages.

Various strategies have been employed to use pre-trained models effectively. For instance, some studies [16, 17] utilize pre-trained encoders like XLSR-53 as feature extractors or encoders paired with a transformer [18] based decoder. Our experiments align with this latter, employing HuBERT variants as the encoder and using a pre-trained Large Language Model (LLMs) transformer based as decoder.

Although the literature does not provide definitive guidance on selecting the most suitable pre-trained speech encoder, [19] observed that encoders trained on the same language as the source language tend to extract more relevant audio features, thereby improving overall performance. Given that Fongbe is an African language and no pre-trained speech model includes it at the time of writing, we hypothesize that using a pre-trained encoder trained on other close linguistically African languages would yield promising results. To test this hypothesis, we conduct training experiments using pre-trained models HUBERT147 and AfriHuBERT as encoders, combined with pre-trained multilingual decoders such as mBART [20] and NLLB [21].

### 3. Fongbe linguistic features

Fongbe, a Gbe language spoken primarily in Benin, serves as a lingua franca for approximately 40–45% of the Beninese population [22]. Fongbe plays a significant role in media, being widely used in both public and private radio and television programs. However, Fongbe tonal nature presents unique challenges, particularly in written and translated texts. In linguistics, tone refers to the use of pitch variations to distinguish meaning in spoken language [23]. Lexical tones, in particular, help differentiate words that are otherwise phonologically identical [24]. These pitch variations, or tonal patterns, are produced by changes in the fundamental frequency of a syllable. For the written form of Fongbe, mainly based on the Latin alphabet with additional symbols, these tones are typically represented using diacritics, which are essential for accurately conveying tonal distinctions in written form. Fongbe primarily features two main tonemes as noted by [25], from which all other tones are derived. In Fongbe, each syllable carries a tone, and the absence of tone marks can often lead to confusion. Regarding tones, there are four tones in Fongbe. The low tone ( ` ), the high tone ( ´ ), the low-high tone ( ˘ ) and lastly, the mid tone ( - ) marked by a small horizontal line. The absence of tone is considered as the mid tone. Fongbe utilizes an alphabet comprising 23 consonants and 12 vowels as shown in Table 1.

Fongbe exhibits a lexicographical structure that is primarily monosyllabic, disyllabic, and trisyllabic shown in table 2. A compatibility study conducted on the combinations of con-

Table 1: *Fongbe consonants and vowels*

Consonants	Vowels
b, c, d, ḍ, f, g, gb, h, j, k, kp, l, m, n, ny, p, r, s, t, v, w, x, y, z	a, e, ε, i, o, ɔ, u

sonants, vowels, and the four tones revealed the presence of 376 monosyllabic structures out of 1,104 meaningful forms. This analysis highlights the phonological richness and structural diversity of Fongbe, underscoring its significance in linguistic studies.

Table 2: *Fongbe syllabic structure*

Structure	Types	Examples
Monosyllabic	V, CV	à, bà
Disyllabic	VCV, CVCV, CVV, VV	azo, galí, fèè, àa
Trisyllabic	VCVCV, CVCVCV, VCVV, CVCVV, CVVCV	asòlò, logosò, agoo, kédédé jaunta

## 4. Data collection process

We augmented the FFSTC corpus [6] by adding new samples selected from a validated set in French, sourced from the Common Voice project [26], a Mozilla Foundation initiative. To reduce the human cost, we utilized the Google Translate to generate Fongbe translations of the French sentences. These translations were then meticulously reviewed and refined by a team of linguists to ensure accuracy and linguistic quality. Once validated, the sentences were uploaded to our custom web application [27] for recording.

Participants, comprising both male and female speakers, were invited to read at least 2,000 sentences each. The reading sessions were conducted in a controlled environment to minimize ambient noise, as Fongbe is a tonal language, and background sounds could interfere with the accurate perception of its tonal distinctions. To further ensure data quality, we carefully selected participants to minimize potential biases arising from regional accents. Specifically, we included only native speakers of Fongbe, excluding individuals who learned Fongbe as a second language or who speak Fongbe with influences from Mahi or Gungbe dialect accents.

The recorded sentences underwent a rigorous validation process by a team of six validators, working in pairs, with each sentence validated once. Sentences containing background noise (e.g., wind or engine noise) or exhibiting incorrect tone patterns were rejected. This meticulous validation process enabled us to successfully add 42,000 new samples to the existing FFSTC corpus.

The FFSTC corpus originally stemmed from a data competition in which multiple participants translated the same French sentences directly into Fongbe. This process resulted in duplicate transcripts and nearly identical speech recordings, contributing to a rich diversity of speech samples. To maximize the potential of this variation, we retained these duplicates in the training set while ensuring that only unique transcripts were included in the validation and test sets. This approach allows future trained models to benefit from the diversity of translations while maintaining data integrity during the evaluation process.

#### 4.1. Statistics of the dataset

As outlined in the introduction, we conducted experiments in both ASR and Speech Translation. For the end-to-end ST task, we utilized the entire dataset. While for the ASR and the cascade ST task, we use only the 36 hours of speech available with their transcripts in Fongbe, as described in Table 3.

Table 3: *Dataset statistics*

Experiments	Split	Hours	Sentences
ASR	Test	3.93	2.5 k
ASR	Valid	3.54	2.4 k
ASR	Train	29	19.9 k
ST	Test	5.9	3.9 k
ST	Valid	6.1	4.1 k
ST	Train	48	29.5 k

## 5. Experiments and results

In this section, we present the experimental framework for (1) ASR system, (2) cascade ST system and end-to-end ST system.

### 5.1. SSL models Description

The use of pre-trained models, as demonstrated in several studies, shows the potential to create efficient recognition or translation systems [28] even with limited amounts of data by fine-tuning them on downstream tasks. For our experiments, we chose to use that method. Among the publicly available pre-trained speech encoders, such as XLSR-128 [29], and HuBERT, we selected HuBERT [30] variants, specifically HuBERT147 and AfriHuBERT, specialized to some African languages (but not to Fongbe). This decision was based on their superior performance on downstream tasks, such as Automatic Speech Recognition (ASR), as demonstrated in [31].

HuBERT is closely related to Wav2Vec 2.0 [32]. While Wav2Vec 2.0 distinguishes between true latent speech representations and the contextualized representations generated by the transformer encoder, HuBERT employs a technique similar to BERT [33] for speech units. Specifically, HuBERT computes a loss over masked speech units, forcing the model to learn high-level representations of unmasked inputs to accurately infer the targets of the masked ones. This approach has been shown to outperform Wav2Vec 2.0 when trained on the same amount of data in [30]. Given these advantages, we expected HuBERT147 and AfriHuBERT to deliver strong performance in our experiments.

mBART is a denoising sequence-to-sequence model pre-trained on high-resource languages. It uses a Transformer architecture to reconstruct texts from noised inputs, where phrases are masked and sentences are permuted. Known for its robustness with noisy data, mBART is particularly well-suited in tasks like speech translation, especially for tonal languages such as Fongbe. NLLB (No Language Left Behind) is a multi-lingual translation model pre-trained on a wide range of languages, including several African languages. Designed for high-quality translation, NLLB aims to bridge the gap between high-resource and low-resource languages, making it a strong candidate for our translation experiments.

### 5.2. ASR experiments

The first experiment was performed using the original Fongbe transcripts, including diacritics, to establish a baseline performance. In the second experiment, we removed the diacritics from the transcripts to evaluate the impact of diacritic removal on recognition accuracy. The third experiment involved a novel approach of diacritic substitution, where we systematically identified monosyllabic words with diacritics and replaced them with their base syllables accompanied by a unique numerical identifier. We report few examples of the substitution on the Table 4. This substitution aimed to modify the representation of diacritics while preserving linguistic information, potentially improving the model’s ability to generalize across similar phonetic patterns.

Diacritic	Replacement	Diacritic	Replacement
ó	o1	ò	o2
õ	o3	ô	o4
á	a1	à	a2
ã	a3	â	a4
é	e1	è	e2
ë	e3	ê	e4

Table 4: *Mapping of diacritics for o, a, and e to numerical representations.*

To conduct the experiments, we trained three different SentencePiece [34] tokenizer models at character level using the combined training and validation sets for each specific case. For the base experiment (with diacritics), the substitution experiment and the experiment without diacritics the vocabulary size are, respectively 62, 44 and 36. This reduction in vocabulary size for the substitution and third experiment case reflects the simplified representation of text when diacritics are either removed or replaced, which in turn may influence the model efficiency and performance. The three ASR models are end-to-end models composed of the AfriHuBERT speech encoder followed by three 1024-dim dense layers. They were fine-tuned on the ASR training dataset by using the CTC loss function. All experiments are run over 50 epochs and results are summarized in the Table 5. ASR recipes will be released for reproducibility.

Table 5: *Word Error Rates (%) on the ASR test dataset*

Experiments	WER
ASR base	21.98
ASR Sub	22.18
ASR without diacritics	17.02

The ASR model trained without diacritics yields the lowest Word Error Rate (WER) of 17.02%, but this WER cannot be compared with the two other ones: the lexical confusion is drastically decreased since removing the diacritics reduce the vocabulary size. Nevertheless, if the automatic transcriptions without diacritics are less informative, these results show they are more reliable. Although the diacritic substitution approach did not outperform the base model, we consider it should be experimented within a cascade speech translation system, because of a different distribution of ASR errors.

### 5.3. Cascade speech translation

Cascade systems for speech translation consist of two key modules: ASR and MT. In our implementation, we used our trained ASR models for the transcription module, followed by an end-to-end text-to-text MT model based on the fine-tuning the NLLB model. Fongbe was included in the NLLB pre-training dataset. We fine-tuned it using the Huggingface [35] trainer. To ensure a fair comparison, we conducted three separate fine-tuning experiments, the first on Fongbe written with diacritics and the second without diacritics and the third on substituted Fongbé. We evaluated the cascade model on the same test set as the end-to-end model presented in the next section.

The fine-tuning of NLLB results yielded BLEU scores of 58.9, 57.56 and 47.39 on manual transcriptions, respectively for the models with diacritics, with diacritics substitution and without diacritics, on the validation subset of ASR containing the Fongbe transcriptions and translations. These results underscore the importance of diacritics in preserving contextual understanding, particularly for tonal languages like Fongbe.

For the experiment using the ASR with the substitution approach, we use a fine-tune the NLLB model on substituted Fongbé text to French. This step ensured that the translation module received input with the correct diacritic markings. The results of the experiments are summarized in Table 6.

Table 6: BLEU scores for the cascade systems on the test dataset

Experiments	BLEU
ASR base + NLLB	32.76
ASR Sub + NLLB	37.23
ASR with diacritics + NLLB	39.60

The best result in the cascade training was achieved by the AfriHuBERT model fine-tuned on ASR with diacritic, reaching a BLEU score of 39.60. Additionally, we observed that the substitution method holds significant potential as it comes in the second position with the BLEU score of 37.23. These experiments reveal that retaining diacritics is more critical for translation of Fongbe than for its recognition, as diacritics provide additional linguistic information about segments that goes beyond what the base syllables alone can convey. However, further studies are needed to fully explore and optimize the substitution approach, as it could provide a viable pathway for improving both efficiency and performance in speech recognition and translation tasks.

### 5.4. End-to-end Speech translation

We conduct several experiments dedicated to the end-to-end approach. We investigated the use of different combinations of the two speech encoders HuBERT-147 and AfriHuBERT with the mBART and NLLB decoders.

All the models were trained on a single V100 32BG GPU with a batch size of 2. We utilized the Adam optimizer and ran the experiments over 50 epochs. To align the output length of the HuBERT encoder with the input dimensions of the mBART and NLLB decoders, we employed a feed-forward layer. During inference, we applied a beam search with a width of 5 to generate translations. To improve our models performance, we applied three data augmentation techniques such as speed perturbation, which slightly speeds up or slows down the audio, frequency drop, which randomly removes certain frequency bands

Table 7: BLEU score of end-to-end speech-to-text translation models

Experiments	Data Aug	Params	BLEU
AfriHuBERT-NLLB	No	962.1M	23.90
AfriHuBERT-NLLB	Yes	-	<b>26.32</b>
AfriHuBERT-mBART	No	553.8M	21.13
AfriHuBERT-mBART	Yes	-	24.30

and chunk drop, which deletes short segments of the audio signal. Since the models based on AfriHuBERT performed better than the models based on HuBERT147, we report in Table 7 only the results reached by using AfriHuBERT as a speech encoder. We notice that AfriHuBERT-NLLB achieved better BLEU scores with data augmentation.

## 6. Conclusion

This work brings a significant advance for Fongbe speech processing and highlights several avenues of future exploration. By extending an existing dataset to 61 hours of high-quality audio and aligned text, we offer the research community a unique and richer resource to build and evaluate speech technologies for Fongbe, a tonal and under-represented language. Our detailed experiments in both cascade and end-to-end Speech Translation reveal several important insights that can stimulate broader research in low-resource language technologies.

Our best cascaded system (automatic speech recognition followed by machine translation) achieved a BLEU score of 39.60. By comparison, the most effective end-to-end model reached a lower, yet still encouraging, BLEU score of 26.32 when enhanced with data augmentation. This performance gap is likely due to the models' handling of tonal information, which is only sparsely represented in the SSL speech encoder. Training an ASR component specifically on this tonal language allows the system to capture tonal cues more accurately, thereby improving downstream translation quality.

The expanded Fongbe corpus and our findings open several possibilities for further research. First, improvements to diacritic substitution, by potentially using more granular markers that capture subtle tonal shifts, could reduce ASR errors while preserving key phonological cues for translation. Second, personalized or speaker-adaptive speech translation models, possibly trained to handle specific dialectal variants, may substantially enhance intelligibility and translation fidelity. Finally, future self-supervised or multilingual pre-training efforts will benefit from explicitly including Fongbe data, leading to more robust encoder-decoder architectures for low-resource African languages.

Overall, this work not only delivers the largest corpus of Fongbe audio currently available for speech recognition and translation, but also highlights data-collection strategies, modeling setups, and diacritic handling approaches that can be generalized to other tonal, underrepresented languages. We release our dataset publicly to support further research<sup>3</sup>.

<sup>3</sup><https://huggingface.co/datasets/GbeBenin/FFSTC-2>

## 7. References

- [1] C. Wang, H. Inaguma, P.-J. Chen, I. Kulikov, Y. Tang, W.-N. Hsu, M. Auli, and J. Pino, "Simple and effective unsupervised speech translation," *arXiv preprint arXiv:2210.10191*, 2022.
- [2] D. Eberhard, G. Simons, and C. Fennig, *Ethnologue: Languages of the World, 24th Edition*, 02 2021.
- [3] I. Adebara and A. . a. Elmadany, "SERENGETI: Massively multilingual language models for Africa." ACL, 2023. [Online]. Available: <https://aclanthology.org/2023.findings-acl.97>
- [4] M. Bala, "© african literature and orality: A reading of ngugi wa thiang'o's wizard of the crow 2007." *JOURNAL OF ENGLISH LANGUAGE AND LITERATURE*, vol. 3, 03 2015.
- [5] F. A. A. LAleye, L. Besacier, E. C. Ezin, and C. Motamed, "First automatic fongbe continuous speech recognition system: Development of acoustic models and language models."
- [6] D. F. Kponou, F. A. A. Laleye, and E. C. Ezin, "FFSTC: Fongbe to French speech translation corpus," in *LREC-COLING 2024*, 2024.
- [7] M. Tachbelie, S. T. Abate, and L. Besacier.
- [8] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof," *LREC*, 2016.
- [9] J. O. Alabi, X. Liu, D. Klakow, and J. Yamagishi, "Afrihubert: A self-supervised speech representation model for african languages," 2024. [Online]. Available: <https://arxiv.org/abs/2409.20201>
- [10] M. Z. Boito and V. I. . al, "mhubert-147: A compact multilingual hubert model," 2024. [Online]. Available: <https://arxiv.org/abs/2406.06371>
- [11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," 2020. [Online]. Available: <https://arxiv.org/abs/2006.13979>
- [12] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.
- [13] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv preprint arXiv:2010.12829*, 2020.
- [14] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pre-training for low-resource speech-to-text translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7909–7913.
- [15] B. B. Odoom, N. Robinson, E. Rippeth, L. Tavarez-Arce, K. Murray, M. Wiesner, P. McNamee, P. Koehn, and K. Duh, "Can synthetic speech improve end-to-end conversational speech translation?" in *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 2024, pp. 167–177.
- [16] J. Mbuya and A. Anastopoulos, "GMU systems for the IWSLT 2023 dialect and low-resource speech translation tasks." [Online]. Available: <https://aclanthology.org/2023.iwslt-1.24>
- [17] M. Zanon Boito and J. . a. Ortega, "ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks." Dublin, Ireland (in-person and online): Association for Computational Linguistics. [Online]. Available: <https://aclanthology.org/2022.iwslt-1.28>
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [19] D. F. Kponou, F. A. A. Laleye, and E. C. Ezin, "Systematic literature review and bibliometric analysis of low-resource speech-to-text translation," K. Arai, Ed. Cham: Springer Nature Switzerland, pp. 379–398.
- [20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," 2020. [Online]. Available: <https://arxiv.org/abs/2001.08210>
- [21] N. Team, M. R. Costa-jussà, and J. C. . al, "No language left behind: Scaling human-centered machine translation," 2022. [Online]. Available: <https://arxiv.org/abs/2207.04672>
- [22] K. J. Gbaguidi, "Taxinomie et analyse des erreurs linguistiques des élèves fonphones en apprentissage de français : Pour une approche linguistique et pragmatique en didactique des langues," Doctoral dissertation, EDP-UAC, 2009.
- [23] B. Caron, "Tone and intonation," in *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*, 2015.
- [24] Y. Xu, "Understanding tone from the perspective of production and perception," *Language and Linguistics*, vol. 5, no. 4, pp. 757–797, 2004.
- [25] C. B. Gnanguènon, "Analyse syntaxique et sémantique de la langue 'fn' au Bénin en Afrique de l'ouest," Ph.D. dissertation, Université Cergy-Pontoise, France, 2014.
- [26] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020. [Online]. Available: <https://arxiv.org/abs/1912.06670>
- [27] K. Fortuné, "Ayihoun.com," 2024, accessed: 2024-12-16. [Online]. Available: <https://ayihoun.com/>
- [28] A. Laurent, S. Gahbiche, H. Nguyen, H. Elleuch, F. Bougares, A. Thiol, H. Riguidel, S. Mdhaffar, G. Laperrière, L. Maison *et al.*, "On-trac consortium systems for the iwslt 2023 dialectal and low-resource speech translation tasks," in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, 2023, pp. 219–226.
- [29] A. Babu, C. Wang, A. Tjandra, and K. L. . al, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," 2021. [Online]. Available: <https://arxiv.org/abs/2111.09296>
- [30] W.-N. Hsu, B. Bolte, and Y.-H. H. T. . al, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021. [Online]. Available: <https://arxiv.org/abs/2106.07447>
- [31] J. O. Alabi, X. Liu, D. Klakow, and J. Yamagishi, "Afrihubert: A self-supervised speech representation model for african languages," 2024. [Online]. Available: <https://arxiv.org/abs/2409.20201>
- [32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [34] T. Kudo, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.
- [35] S. M. Jain, "Hugging face." Springer, 2022, pp. 51–67.