FFSTC 2: Extending the Fongbe to French Speech Translation Corpus

D. Fortuné Kponou
IMSP Dangbo Bénin
fortune.kponou@msp-uac.org

Salima Mdhaffar
LIA Avignon France
salima.mdhaffar@univ-avignon.fr

Fréjus A. A. Laleye
OPSCIDIA Paris France
frejus.laleye@opscidia.com

Eugène C. Ezin
IMSP Dangbo Bénin
eugene.ezin@imsp-uac.org

Yannick Estève
LIA Avignon France
yannick.esteve@univ-avignon.fr

Abstract

This paper introduced 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 this enriched corpus, we developed both cascade and end-to-end speech translation systems. Our models employ AfriHuBERT and HuBERT147, two speech encoders specialized to African languages, and the NLLB and mBART models as decoders. We also investigate the use of the SAMU-XLSR approach to inject sentence-level semantic information to the XSLR-128 model used as an alternative speech encoder. We also introduced a novel diacritic-substitution technique for ASR, which, when combined with NLLB, enables a cascade model to achieve a BLEU score of 37.23 ompared to 39.60 obtained by the best system using original diacritics. Among the end-to-end architectures evaluated, the architectures with data augmentation and NLLB as decoder achieved the highest score respectively, SAMU-NLLB scored the BLEU score of 28.43.

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 (Wang et al., 2022). 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 (Eberhard et al., 2021), yet, as highlighted by (Adebara and Elmadany, 2023), 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[1] 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 (Bala, 2015) 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 (Laleye et al.) dataset for speech recognition, which includes 6 hours of audio along with transcriptions, and the FFSTC dataset (Kponou et al., 2024), which contains 31 hours of Fongbe speech paired with French trans-

[1] https://commonvoice.mozilla.org

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 145–152

lations. Also, corpora like GigaST are pseudo-labeled, distinguishing them from fully human-annotated datasets such as (Tachbelie et al.; Gauthier et al., 2016) 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 (Kponou et al., 2024). 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 (Alabi et al., 2024; Boito and al, 2024), 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[2]. 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 (Conneau et al., 2020), which is pre-trained on 53 languages data. Studies (Bansal et al., 2018; Li et al., 2020; Stoian et al., 2020) 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, 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 (Odoom et al., 2024). 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 (Boito and al, 2024), which supports 16 African languages, and its variant AfriHuBERT (Alabi et al., 2024), which extends coverage to 39 languages.

Various strategies have been employed to use pre-trained models effectively. For instance, some studies (Mbuya and Anastasopoulos; Zanon Boito and Ortega) utilize pre-trained encoders like XLSR-53 as feature extractors or encoders paired with a transformer (Vaswani et al., 2023) 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, (Kponou et al.) 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 (Liu et al., 2020) and NLLB (Team et al., 2022).

3 Fongbe linguistic features

Fongbe, a Gbe language spoken primarily in Benin, serves as a lingua franca for approximately $40\tilde{}45\%$ of the Beninese population (Gbaguidi, 2009). Fongbe plays a significant role in media, being widely used in both public and private radio

[2]International Workshop on Spoken Language Translation

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 (Caron, 2015). Lexical tones, in particular, help differentiate words that are otherwise phonologically identical (Xu, 2004). 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 (Gnanguènon, 2014), 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.

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

Table 1: Fongbe consonants and vowels

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

Structure	Types	Examples
Monosyl.	V, CV	à, bà
Disyllabic	VCV, CVCV, CVV, VV	azo, galí, fɛ̀è, àa
Trisyllabic	VCVCV, CVCVCV, VCVV, CVCVV CVVCV	asòlò, logosò, agoo, kɛ́dɛ́ɛ́ jaunta

Table 2: Fongbe syllabic structure

4 Data collection process

We augmented the FFSTC corpus (Kponou et al., 2024) by adding new samples selected from a validated set in French, sourced from the Common Voice project (Ardila et al., 2020), 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 (Fortuné, 2024) 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 Dataset statistics

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.

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

Table 3: Dataset statistics

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 (Laurent et al., 2023) 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 (Babu et al., 2021), and HuBERT, we selected HuBERT (Hsu et al., 2021) 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 (Alabi et al., 2024).

HuBERT is closely related to Wav2Vec 2.0 (Baevski et al., 2020). While Wav2Vec 2.0 distinguishes between true latent speech representations and contextualized representations generated by the transformer encoder, HuBERT employs a technique similar to BERT (Devlin et al., 2019) 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 (Hsu et al., 2021). Given these advantages, we expected HuBERT147 and AfriHuBERT to deliver strong performance in our experiments.

We also trained a SAMU_XLSR model using our dataset. Unlike the approach in (Khurana et al., 2022), which relies on speech transcripts for alignment, we used translated labels instead. SAMU is built on XLSR and utilizes a frozen Language-Agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2020) as the master model to semantically align Fongbe speech and French text embeddings in the XLSR space. We trained SAMU for 50 epochs on the ST training dataset.

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 multilingual 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. 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 as reported in Table 4.

To conduct the experiments, we trained three different SentencePiece (Kudo, 2018) 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

State	Sentence
Diacritics	tavo ayihun tɔn dé dò disixwé **transl.**: *a game table on the right*
w/o diacr.	tavo ayihun tɔn de do disixwe
Substitution	tavo ayihun tɔn de1 do2 disixwe1

Table 4: Example of diacritic processing

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.

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

Table 5: Word Error Rates (%) on the ASR test dataset reached by the AfriHuBERT speech encoder (*our best results*)

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 (Machine Translation). 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 (Jain, 2022) trainer. To ensure a fair comparison, we conducted three separate fine-tuning experiments, the first on Fongbe written with diacritics, the second with substituted Fongbe and the last on Fongbe without diacritics. 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 and without diacritics, with substitution and without diacritics, on the validation subset containing the Fongbe transcriptions. 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 fine-tune the model NLLB using a substituted Fongbe. This step ensured that the translation module will receive the correct input for each case. The results of the experiments are summarized in Table 6.

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

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

The best result in cascade training was achieved by the AfriHuBERT model fine-tuned on ASR with diacritic, reaching a BLEU score of 39.60 follow by the susbstitution system (ASR Sub) with the BLEU 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. Additionally, we observed that the substitution method holds significant potential. However, further studies are needed to fully explore and optimize this 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 conducted several experiments dedicated to the end-to-end approach. We investigated the use of different speech encoders HuBERT-147 and AfriHuBERT with different decoders mBART and NLLB. We combined different augmentations to perform data augmentation: Speed perturbation (resample the audio signal at a rate that is similar to the original rate, to achieve a slightly slower or slightly faster signal), Frequency drop (randomly drops a number of frequency bands to zero) and Chunk drop (Chunk drop is an augmentation strategy helps a models learn to rely on all parts of the signal, since it can't expect a given part to be present).

Experiments	Aug	Params	**BLEU**
AfriHuBERT-NLLB	No	962.1M	23.90
AfriHuBERT-NLLB	Yes	962.1M	26.32
AfriHuBERT-mBART	No	553.8M	22.16
AfriHuBERT-mBART	Yes	553.8M	24.30
SAMU-mBART	No	1.4B	25.11
SAMU-mBART	Yes	1.4B	24.17
SAMU-NLLB	No	1.8B	25.85
SAMU-NLLB	Yes	1.8B	28.43

Table 7: BLEU score of end-to-end speech-to-text translation models, with of without data augmentation.

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 enhance their performance, we applied data augmentation to each model. 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 observed that SAMU achieved better BLEU scores with data augmentation than the other end-to-end, as documented in Table 7. We conclude that semantic alignment in the embedding space of SAMU provides it with a better speech representation for the decoder.

NLLB's broader linguistic coverage did not translate into superior performance in our experiments.

6 Conclusion

This work represents a significant advance for Fongbe speech processing, for both transcription and translation to French. 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 cascade system achieved a BLEU score of 39.60, underscoring the power of carefully handling tonal information. In contrast, our most effective end-to-end model achieved a BLEU of 28.43, especially when leveraging data augmentation and semantic alignment.

The expanded Fongbe corpus and our findings open several possibilities for further research. First, improvements to diacritic substitution—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, modelling setups, and diacritic handling approaches that can be generalized to other tonal, under-represented languages.

References

Ife Adebara and AbdelRahim al Elmadany. 2023. SERENGETI: Massively multilingual language models for Africa. ACL.

Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, and Junichi Yamagishi. 2024. Afrihubert: A self-supervised speech representation model for african languages. *Preprint*, arXiv:2409.20201.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers,

and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Arun Babu, Changhan Wang, Andros Tjandra, and Kushal Lakhotia al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Mustapha Bala. 2015. © african literature and orality: A reading of ngugi wa thiang'o's wizard of the crow 2007. *JOURNAL OF ENGLISH LANGUAGE AND LITERATURE*, 3.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Marcely Zanon Boito and Vivek Iyer al. 2024. mhubert-147: a compact multilingual hubert model. *Preprint*, arXiv:2406.06371.

Bernard Caron. 2015. Tone and intonation. In *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *Preprint*, arXiv:2006.13979.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

David Eberhard, Gary Simons, and Chuck Fennig. 2021. *Ethnologue: Languages of the World, 24th Edition*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Kponou Fortuné. 2024. Ayihoun.com. Accessed: 2024-12-16.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. *LREC*.

K. J. Gbaguidi. 2009. *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.

C. B. Gnanguènon. 2014. *Analyse syntaxique et sémantique de la langue "fn" au Bénin en Afrique de l'Ouest*. Ph.D. thesis, Université Cergy-Pontoise, France.

Wei-Ning Hsu, Benjamin Bolte, and Yao-Hung Hubert Tsai al. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Shashank Mohan Jain. 2022. Hugging face. pages 51–67. Springer.

Sameer Khurana, Antoine Laurent, and James Glass. 2022. Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1493–1504.

D. Fortuné Kponou, Fréjus A. A. Laleye, and Eugène C. Ezin. Systematic literature review and bibliometric analysis of low-resource speech-to-text translation. pages 379–398, Cham. Springer Nature Switzerland.

D. Fortuné Kponou, Fréjus A. A. Laleye, and Eugène Cokou Ezin. 2024. FFSTC: Fongbe to French speech translation corpus. In *LREC-COLING 2024*.

T Kudo. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Fréjus A. A. Laleye, Laurent Besacier, Eugène C. Ezin, and Cina Motamed. First automatic fongbe continuous speech recognition system: Development of acoustic models and language models.

Antoine Laurent, Souhir Gahbiche, Ha Nguyen, Haroun Elleuch, Fethi Bougares, Antoine Thiol, Hugo Riguidel, Salima Mdhaffar, Gaëlle Laperrière, Lucas Maison, and 1 others. 2023. 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)*, pages 219–226.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Jonathan Mbuya and Antonios Anastasopoulos. GMU systems for the IWSLT 2023 dialect and low-resource speech translation tasks.

Bismarck Bamfo Odoom, Nathaniel Robinson, Elijah Rippeth, Luis Tavarez-Arce, Kenton Murray, Matthew Wiesner, Paul McNamee, Philipp Koehn,

and Kevin Duh. 2024. 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)*, pages 167–177.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Martha Tachbelie, Solomon Teferra Abate, and Laurent Besacier. Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic.

NLLB Team, Marta R. Costa-jussà, and James Cross al. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2022. Simple and effective unsupervised speech translation. *arXiv preprint arXiv:2210.10191*.

Yi Xu. 2004. Understanding tone from the perspective of production and perception. *Language and Linguistics*, 5(4):757–797.

Marcely Zanon Boito and John al 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.