FFSTC: Fongbe to French Speech Translation Corpus

D. Fortuné Kponou[1], Fréjus A. A. Laleye[2], Eugène C. Ezin[1]

1 Institut de Mathématiques et de Sciences Physiques, Dangbo, Bénin

2 OPSCIDIA, Paris, France

{fortune.kponou,eugene.ezin}@imsp-uac.org,

frejus.laleye@opscidia.com

Abstract

In this paper, we introduce the Fongbe to French Speech Translation Corpus (FFSTC). This corpus encompasses approximately 31 hours of collected Fongbe language content, featuring both French transcriptions and corresponding Fongbe voice recordings. FFSTC represents a comprehensive dataset compiled through various collection methods and the efforts of dedicated individuals. Furthermore, we conduct baseline experiments using Fairseq's transformer_s and conformer models to evaluate data quality and validity. Our results indicate a score BLEU of 8.96 for the transformer_s model and 8.14 for the conformer model, establishing a baseline for the FFSTC corpus.

Keywords: Speech translation corpus, spoken language translation, low-resource language, Fongbe-French, Fongbe

1. Introduction

In the era of global communication and rapid technological advancement, the development of efficient translation systems holds immense importance. These communications systems have the potential to break language barriers, facilitating meaningful knowledge exchange, and increasing the cross-cultural connections among individuals from diverse linguistic backgrounds (Doherty, 2016). They play a pivotal role in enhancing activities such as tourism, contributing to economic growth (Lenba and Ennebati, 2022). Translation systems help people to access educational resources and information in their native languages, promoting literacy and knowledge sharing.

While text-based translation systems have revolutionized communication for many languages, they fall short for tonal languages like Fongbe, which are more spoken than written. This creates a critical need for speech-to-text translation tailored to tonal languages such as Fongbe. Fongbe is the most spoken dialect of Benin, by more than 50% of Benin's population, including 8 million speakers. Fongbe is also spoken in Nigeria and Togo (Laleye et al., 2016).

Speech translation involves the transformation of spoken language into text in another language. It is worth noting that speech translation has predominantly focused on languages with abundant linguistic resources, including English, French, Chinese, and Spanish (Qi et al., 2022; Jia et al., 2022; Iranzo-Sánchez et al., 2020). The traditional speech translation approach called cascade method, involve two separate modules, the first module for automatic speech recognition (ASR) and the second for text translation (Etchegoyhen et al., 2022). This approach require two separate corpora to train individually each modules that are subsequently coupled. The advent of sequence-to-sequence architectures (Sutskever et al., 2014) has reshaped this landscape, enabling the creation of a single corpus containing audio recordings and target texts. This development is particularly relevant for low-resource languages like Fongbe, which are in the category of spoken language and have limited written resources. So it is challenging to get written text in Fongbe and its corresponding in French, than getting a recording of voice in Fongbe with transcription in French.

In response to these challenges and opportunities, we present the Fongbe-French Speech Translation Corpus (FFSTC), a unique resource comprising 31 hours of spoken Fongbe paired with French text. Our objective is to advance voice translation technologies for less commonly studied languages like Fongbe.

This paper makes two key contributions. The first is the introduction of the Fongbe-French Speech Translation Corpus (FFSTC), the first public dataset of its kind. This rich dataset, contain 31 hours of speech data with diverse speakers and topics. The second is the establishment of a baseline performance by evaluating a transformer_s and conformer model of Fairseq toolkit (Ott et al., 2019) on the FFSTC, achieving a baseline BLEU of 8.96 and 8.14 respectively. This benchmark provides a reference point for future research using the FFSTC and helps identify areas for improvement.

This paper is structured as follows. Section 2 provides an overview of existing related studies. Section 3 delves into the methodology used to create the FFSTC dataset, detailing the data collection and processing procedures. Section 4 then explores the structure and statistics of the corpus, offering

insights into its composition and characteristics. Section 5 presents the experiments conducted to establish a baseline performance for a transformer and conformer model on the FFSTC, using BLEU metrics to assess its effectiveness. Finally, Section 6 concludes the paper by summarizing the key contributions and outlining potential future directions.

2. Related work

Large-scale speech translation datasets play a crucial role in advancing research in speech translation. However, the availability of such datasets is limited, especially for low-resource languages. To achieve meaningful breakthroughs in speech translation, it is crucial to acknowledge that, even though the research in this domain is relatively recent, substantial efforts must be focused on the construction of parallel datasets. Recent available datasets encompass a variety of language pairs, including Chinese-Mongolian (Qi et al., 2022), Chinese-English (Zhang et al., 2021), translations from 21 languages into English (Jia et al., 2022), and translations from English into 15 languages (Wang et al., 2020c). The introduced corpora, such as GigaST (Ye et al., 2022), LibriVox (Beilharz et al., 2020), LibriSpeech-FR (Kocabiyikoglu et al., 2018), mintzai-ST (Etchegoyhen et al., 2021), Multilingual TEDx Corpus (Salesky et al., 2021), and Europarl-ST (Iranzo-Sánchez et al., 2020), contribute to filling the gap of limited resources available for training end-to-end speech translation models. These datasets enable researchers to develop more robust and effective systems for various well-resourced languages. It is therefore undeniable that speech translation research has been extensively conducted for major languages like English, Japanese, and Spanish. However, there is a lack of research in speech translation for under-resourced languages. Some efforts have nevertheless been made for the creation of resources for certain low-resource languages. Woldeyohannis et al. (2018), have created an Amharic speech corpus by preparing 7.43 hours of read-speech from $8,112$ sentences. Additionally, they have developed a parallel Amharic-English corpus of $19,972$ sentences with tourism as the application domain. Another study focuses on constructing a large corpus for speech translation from Khmer (Cambodian) to English and French (Soky et al., 2021). The corpus includes approximately 155 hours of speech and 1.7 million words of text from the Extraordinary Chambers in the Courts of Cambodia (ECCC). Moreover, another project aims to provide datasets for Tamasheq, a developing language spoken in Mali and Niger (Boito et al., 2022). The datasets consist of radio recordings from Studio Kalangou in Niger and Studio Tamani in Mali. They include a large amount of unlabeled audio data in five languages (French, Fulfulde, Hausa, Tamasheq, and Zarma) and a smaller 17 hour parallel corpus of audio recordings with translations in French. The quantity of data collected for the Ahmaric and Tamasheq languages, compared to other well-resourced languages, confirms the difficulty in establishing solid dataset for a system of speech translation for low-resourced languages. For these languages, the lack of available speech and text corpora remains a significant challenge for speech translation.

3. Methodology

In this section, we delve into the intricate process of creating the corpus, shedding light on the meticulous steps taken to acquire and compile the translations. Our methodology not only outlines the data collection procedure but also discusses the quality control measures implemented to ensure the reliability of the dataset.

3.1. Clips creation process

The data within this corpus originates from three distinct sources, each contributing significantly to its composition. The first source is the ALFFA corpus (Laleye et al., 2016), from which we extracted sentences in the Fongbe language and corresponding audio recording. These sentences were subsequently translated into French by experienced linguists.

The second source is the FFR1.1 corpus (Dossou and Emezue, 2020), which serves as a corpus for machine translation. It comprises sentences in French, accompanied by their written translations in the Fongbe language. In this case, we provided the Fongbe sentences to linguists who read and recorded them through a dedicated web platform.

Finally, the third source involves an assembly of sentences gathered from various books, including "Kondo le requin" (Pliya, 1981) and "Un piège sans fin" (Olympe, 1960). We also excerpts from texts found on news websites. These sentences were thoughtfully selected, condensed, and made accessible via a web platform. Given the complexity of this task, we encouraged participants to collaborate within teams to ensure the coherence and quality of the collection process.

The participants teams provided Fongbe translations for the given sentences in French. They committed to a voice recording process using laptops or smartphones through our web platform's interface. This involved audibly translating the sentences displayed in French on the screen. Following this, recorded submissions underwent a rigorous validation process by designated validators. These validators used a straightforward voting system to assess and determine the quality of each translation.

3.2. Clips validation process

In the case of the first source, French transcripts derived from Fongbe transcriptions undergo a meticulous peer-review process by a second validator. This stringent review ensures the translation's accuracy and conformity.

For the second source, the validation process adopts a straightforward binary approach, utilizing a yes-or-no voting system. Audio recordings that align precisely with the provided Fongbe transcripts are retained, while those that do not are omitted.

The third source follows a selective inclusion process. Only clips endorsed as valid by the system are integrated into the corpus. Evaluation criteria encompass multiple facets, including translation completeness, audio recording clarity, and overall translation quality. Our platform features a robust voting system that facilitates the filtration of clips created by participants based on validators' assessments. The voting system employs a scale ranging from zero to five. Recordings that receive a score below two are classified as invalid and are consequently excluded from the dataset. Conversely, recordings scoring three or higher are regarded as valid contributions and are included in the dataset for further analysis. Subsequently, all recordings undergo a comprehensive review by a dedicated team of validators, who make the final determination of approval or rejection.

4. Corpus structure and statistics

In the previous section of the report, we outlined the origins of our dataset, which emanate from three distinct sources. The ASR corpus ALFFA (Laleye et al., 2016) translation, a recording selection of sentences extracted from the FFR1.1 corpus (Dossou and Emezue, 2020) with recorded readings, and a collaborative effort within a team for translating French sentences.

Combining the first two sources, we accumulated a total of 10 hours of data, while the latter made a significant contribution of 21 hours. The team work was done by a group of volunteers who were very committed. There were 8 women and 12 men in the group, all between 20 and 40 years old, making a total of 20 people.

Following the corpus collection, we splited it into training (Train), development (Dev), and test (Test) sets, as outlined in Table 1. The corpus, in its entirety, comprises $16,447$ sentences, which were presented randomly to the participants. Additionally, multiple participants within each team were allowed to record simultaneously using the same credentials.

Figure 1 illustrates the distribution of vocabulary and word counts across the Train, Dev, and Test partitions of our collected dataset for the Fongbe speech translation task. Figure 1 highlights the sub-

stantial vocabulary diversity and word richness in our dataset, with Train exhibiting the highest vocabulary count at approximately $7,500$ unique words. This variation in vocabulary and word counts across splits underscores the dataset's suitability for training and evaluating speech translation models with varying language complexities.
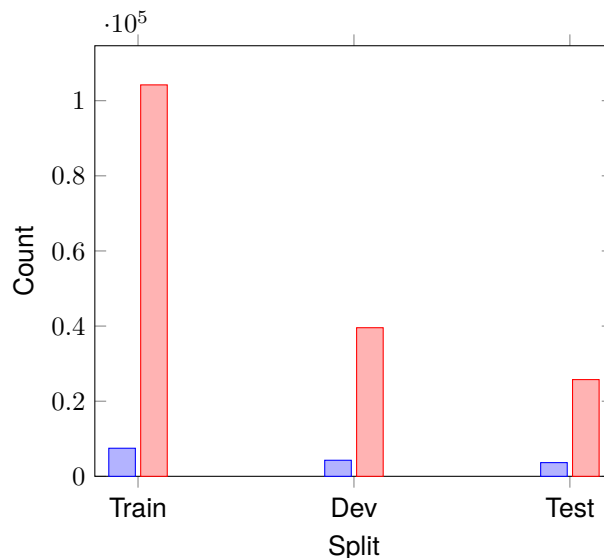

Figure 1: Vocabulary and number of words of each split

Split	Sentences	Audio (hours)
Train	11636	20
Dev	2329	7
Test	2482	4
Minimum sentence length (words)		
Train	1	-
Dev	1	-
Test	1	-
Maximum sentence length (words)		
Train	136	-
Dev	45	-
Test	63	-

Table 1: Corpus statistics.

5. Fongbe-French End-to-End Speech Translation Baseline

In this work, we use the Fairseq toolkit to establish a baseline for Fongbe-French End-to-End speech translation. We run the baseline experiments using a state-of-the-art approach, implementing the transformer_s and conformer architecture within Fairseq framework to build a speech translation model.

5.1. Experiments

We ran the experiments on the collected dataset with the downsampling of the initially recorded au-

dio clips from 44.1 kHz to 16 kHz during the preprocessing step. Following this, to maintain uniformity and streamline the training process, we subjected the extracted features to a normalization procedure, aligning them to a mean of 0 and a standard deviation of 1. As part of our data quality control measures, we implemented a filtering step to exclude samples that exceeded $3,000$ frames, thus ensuring the quality and manageability of the dataset.

We used for 80-dimensional mel-filter bank features and built character-level vocabulary using Sentencepiece ([Kudo and Richardson, 2018](#)). We employed a label-smoothed cross-entropy loss function with a label-smoothing factor set to 0.1. Optimizing the models utilized the Adam optimizer with a learning rate of $1e-3$, coupled with a learning rate scheduler of the inverse square root type, featuring warm-up updates and gradient clipping. The model is trained for 500 epochs using the Adam optimizer ([Kingma and Ba, 2014](#)).

5.2. Results and discussion

We utilized a beam size of 5 and averaged the last 10 checkpoints when evaluating the baseline model trained on the test set. Our evaluation metric was case-insensitive tokenized BLEU, implemented using the sacreBLEU toolkit ([Post, 2018](#)) following the methodology proposed by ([Papineni et al., 2002a](#)).

Architecture	enc-type	FFSTC (test)
transformer_s	-	8.96 (18.4/10.9/10.7/10.7)
conformer	abs	8.14 (16.3/8.8/8.5/8.4)

Table 2: Baseline for Fongbe-French End-to-end speech translation BLEU-4 results.

The BLEU score achieved on the test set of the collected dataset using the transformer_s and conformer architecture, is reported in Table 2. The reported score is dissected into precision scores for various n-gram lengths. For the transformer_s there is 28.7% for unigrams (1-grams), 4.6% for bigrams (2-grams), 1.8% for trigrams (3-grams), and 0.6% for 4-grams. For the conformer there is 16.3 for 1-grams, 8.8 for 2-grams, 8.5 for 3-grams and 8.4 for 4-grams. We set the positional encoding for the conformer to absolute. This evaluation provides valuable insights into the quality of our collected dataset. Despite the relatively low BLEU score of 8.96 for the transformer_s and 8.14 for the conformer, which signifies room for improvement, it is worth noting that this score holds promise for the development of an effective Fongbe-French end-to-end speech translation model. While there is certainly room for enhancement, particularly in terms of n-gram precision, this result serves as

a foundation upon which further refinements and optimizations can be built.

In future research endeavors, we anticipate that the integration of pre-training models and advanced techniques, such as transfer learning, could significantly enhance the performance of the baseline model. Leveraging the rich resources of pre-training models in French and further optimization strategies promises to elevate the accuracy and fluency of Fongbe-French speech translation, thus bridging linguistic and cultural divides more effectively. This work lays the groundwork for future studies aimed at harnessing the full potential of state-of-the-art models and data-driven approaches to advance the field of speech translation for underrepresented languages like Fongbe.

6. Conclusion

In this paper, we introduced FFSTC, the first speech translation corpus for Fongbe language. This dataset has been carefully curated, including a wide variety of language details and difficulties for tonal language. The corpus is valuable for researchers and professionals who are working on speech translation and natural language processing. It stands as the first of its kind for the West African languages, contributing not only to the advancement of Fongbe-French translation but also to the broader goal of preserving and promoting linguistic diversity. Our baseline experiments, conducted on this novel resource, demonstrate promising potential for the development of robust speech translation systems. By providing a dataset focused on a tonal language, the purpose is to explore new methodologies, and potentially develop breakthrough technologies that could benefit not only Fongbe speakers but speakers of other tonal and underrepresented languages. We envision that this dataset will play a pivotal role in bridging linguistic and cultural divides, facilitating enhanced communication and mutual understanding across languages and communities in the West African context. Its availability paves the way for exciting future research endeavors and applications in the field.

7. Acknowledgements

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang.

2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*.

Jeong-Uk Bang, Joon-Gyu Maeng, Jun Park, Seung Yun, and Sang-Hun Kim. 2023. English–korean speech translation corpus (enkost-c): Construction procedure and evaluation results. *ETRI Journal*, 45(1):18–27.

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. LibriVoxDeEn: A corpus for German-to-English speech translation and German speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3590–3594, Marseille, France. European Language Resources Association.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estéve. 2022. Speech resources in the tamasheq language. *Language Resources and Evaluation Conference (LREC)*.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. 2021. Kosp2e: Korean speech to english translation corpus. *arXiv preprint arXiv:2107.02875*.

B.B. Dadié. 1955. *Le pagne noir: contes africains*. Présence africaine: Ed. Africaines. Présence africaine.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

Stephen Doherty. 2016. The impact of translation technologies on the process and product of translation. *International Journal of Communication*, 10:969.

Bonaventure FP Dossou and Chris C Emezue. 2020. Ffr v1. 1: Fon-french neural machine translation. *arXiv preprint arXiv:2006.09217*.

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete, Aitor Alvarez, Iván G. Torre, Juan Manuel Martín-Doñas, Ander González-Docasal, and Edson Benites Fernandez. 2022. Cascade or direct speech translation? a case study. *Applied Sciences*, 12(3).

Thierry Etchegoyhen, Haritz Arzelus, Harritxu Gete Ugarte, Aitor Alvarez, Ander González-Docasal, and Edson Benites Fernandez. 2021. mintzai-ST: Corpus and Baselines for Basque-Spanish Speech Translation. In *Proc. IberSPEECH 2021*, pages 190–194.

Christian Federmann and William D. Lewis. 2016. Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Taku Kudo and John Richardson. 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. 2016. First automatic fongbe continuous speech recognition system:

Development of acoustic models and language models. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 477–482.

Noureddine Lenba and Fatima Zohra Ennebati. 2022. The role of translation in the development of tourism in algeria : Obstacles and opportunities.

Linh The Nguyen, Nguyen Luong Tran, Long Doan, Manh Luong, and Dat Quoc Nguyen. 2022. A high-quality and large-scale dataset for english-vietnamese speech translation.

BHÊLY-QUENUM Olympe. 1960. *Un piège sans fin*, editions présence africaine edition.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jean Pliya. 1981. *Kondo, le requin*, editions clé en coédition avec nena edition.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Xiao Ke Qi, Borjigin B.Teniger, Yuan Sun, and Xiaobing Zhao. 2022. A dataset of Mongolian-Chinese speech translation.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *ArXiv*, abs/2102.01757.

Kak Soky, Masato Mimura, Tatsuya Kawahara, Sheng Li, Chenchen Ding, Chenhui Chu, and Sethserey Sam. 2021. Khmer speech translation corpus of the extraordinary chambers in the courts of cambodia (eccc). In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 122–127.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020b. Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Changhan Wang, Anne Wu, and Juan Pino. 2020c. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Michael Melese Woldeyohannis, Laurent Besacier, and Million Meshesha. 2018. A corpus for amharic-english speech translation: The case of tourism domain. In *Information and Communication Technology for Development for Africa*, pages 129–139, Cham. Springer International Publishing.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2022. Gigast: A 10,000-hour pseudo speech translation corpus. *arXiv preprint arXiv:2204.03939*.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. Bstc: A large-scale chinese-english speech translation dataset.