IMS' Systems for the IWSLT 2021 Low-Resource Speech Translation Task

Pavel Denisov, Manuel Mager, Ngoc Thang Vu
Institute for Natural Language Processing, University of Stuttgart
{pavel.denisov,manuel.mager,thangvu}@ims.uni-stuttgart.de

Abstract

This paper describes the submission to the IWSLT 2021 Low-Resource Speech Translation Shared Task by IMS team. We utilize state-of-the-art models combined with several data augmentation, multi-task and transfer learning approaches for the automatic speech recognition (ASR) and machine translation (MT) steps of our cascaded system. Moreover, we also explore the feasibility of a full end-to-end speech translation (ST) model in the case of very constrained amount of ground truth labeled data. Our best system achieves the best performance among all submitted systems for Congolese Swahili to English and French with BLEU scores 7.7 and 13.7 respectively, and the second best result for Coastal Swahili to English with BLEU score 14.9.

1 Introduction

We participate in the low-resource speech translation task of IWSLT 2021. This task is organized for the first time, and it focuses on three speech translation directions this year: Coastal Swahili to English (swa→eng), Congolese Swahili to French (swc→fra) and Congolese Swahili to English (swc→eng). Working on under-represented and low-resource languages is of special relevance for the inclusion into technologies of big parts of the world population. The Masakhane initiative (Nekoto et al., 2020) has opened the doors for large scale participatory research on languages of the African continent, to which Swahili belongs to. Our Speech-to-Text translation systems aim to contribute to this global effort.

A common problem for these languages is the small amount of data. This is also true for the language pairs of the shared task: the provided data contains a small amount of translated speech samples for each pair, but the participants are allowed to use additional data and pre-trained models for the sub-tasks of ASR and MT. We utilize most of the suggested additional data resources to train and tune sequence-to-sequence ASR and MT components. Our primary submission is the cascaded system built of Conformer end-to-end ASR model and Transformer MT model. Our contrastive system is end-to-end ST system utilizing parameters transfer from the Encoder part of ASR model and the full MT model.

Both ASR and MT components of the cascaded system initially yield good results on their own, but the discrepancy between language formats (spoken vs. written) in ASR and MT corpora causes degradation by 47% in resulting scores. To adapt the MT system to the output of the ASR, we transform the Swahili source data to output similar to one of an ASR system. To further increase the performance of our MT system, we leverage both source formats (original Swahili text and simulated ASR output Swahili) into a multi-task framework. This approach improves our results by 17%, mostly for the English target language. Our system outperforms the next best system on swc→fra by 4.4 BLEU points, but got outperformed by 10.4 BLEU for swa→eng, being the second-best team. Our team was the only participating for swc→eng language pair with a score of 7.7 BLEU. The results of end-to-end system consistently appear to be about twice worse compared to the pipeline approach.

2 ASR

2.1 Data

Table 1 summarizes the datasets used to develop our ASR system. The training data comprises of the shared task training data, Gamayun Swahili speech samples[1] and the training subsets of ALFFA dataset (Gelas et al., 2012) and IARPA Babel Swahili Lan-

[1] https://gamayun.translatorswb.org/data/

Proceedings of the 18th International Conference on Spoken Language Translation, pages 175–181
Bangkok, Thailand (Online), August 5–6, 2021. ©2021 Association for Computational Linguistics

guage Pack (Andresen et al., 2017). The validation data comprises of 869 randomly sampled utterances from the shared task training data and the testing subset of ALFFA dataset. The testing data is the shared task's validation data. All audio is converted to 16 kHz sampling rate. Applied data augmentation methods are speed perturbation with the factors of 0.9, 1.0 and 1.1, as well as SpecAugment (Park et al., 2019). Transcriptions of the shared task data and Gamayun Swahili speech samples dataset are converted from written to spoken language similarly to Bahar et al. (2020), namely all numbers are converted to words[2], punctuation is removed and letters are converted to lower case. External LM is trained on the combination of transcriptions of the ASR training data and LM training data from ALFFA dataset. The validation data for the external LM contains only transcriptions of the ASR validation data.

Dataset	Training		Validation		Testing	
	Utt.	Hours	Utt.	Hours	Utt.	Hours
IWSLT'21 swa	4,162	5.3	434	0.5	868	3.7
IWSLT'21 swc	4,565	11.1	435	1.0	868	4.2
Gamayun	4,256	5.4	-	-	-	-
IARPA Babel	21,891	28.8	-	-	-	-
ALFFA	9,854	9.5	1,991	1.8	-	-
Total	44,728	60.3	2,860	3.4	1,736	7.9

Table 1: Datasets used for the ASR system.

2.2 Model

The ASR system is based on end-to-end Conformer ASR (Gulati et al., 2020) and its ESPnet implementation (Guo et al., 2020). Following the latest LibriSpeech recipe (Kamo, 2021), our model has 12 Conformer blocks in Encoder and 6 Transformer blocks in Decoder with 8 heads and attention dimension of 512. The input features are 80 dimensional log Mel filterbanks. The output units are 100 byte-pair-encoding (BPE) tokens (Sennrich et al., 2016). The warm-up learning rate strategy (Vaswani et al., 2017) is used, while the learning rate coefficient is set to 0.005 and the number of warm-up steps is set to 10000. The model is optimized to jointly minimize cross-entropy and connectionist temporal classification (CTC) (Graves et al., 2006) loss functions, both with the coefficient of 0.5. The training is performed for 35 epochs on 2 GPUs with the total batch size of 20M bins and gradient accumulation over each 2 steps. After that, 10 checkpoints with the best validation accuracy are averaged for the decoding. The decoding is performed using beam search with the beam size of 8 on the combination of Decoder attention and CTC prefix scores (Kim et al., 2017) also with the coefficients of 0.5 for both. In addition to that, external BPE token-level language model (LM) is used during the decoding in the final ASR system. The external LM has 16 Transformer blocks with 8 heads and attention dimension of 512. It is trained for 30 epochs on 4 GPUs with the total batch size of 5M bins, the learning rate coefficient 0.001 and 25000 warm-up steps. Single checkpoint having the best validation perplexity is used for the decoding.

2.3 Pre-trained models

In addition to training from scratch, we attempt to fine-tune several pre-trained speech models. These models include ESPnet2 Conformer ASR models from the LibriSpeech (Panayotov et al., 2015), SPGISpeech (O'Neill et al., 2021) and Russian Open STT[3] recipes, as well as wav2vec 2.0 (Baevski et al., 2020) based models XLSR-53 (Conneau et al., 2020) and VoxPopuli (Wang et al., 2021).

2.4 Results

Table 2 summarizes the explored ASR settings and the results on the shared task validation data. CTC weight 0.5 is selected in order to minimize the gap between ASR accuracy on the two Swahili languages. Evaluation of pre-trained English ASR models expectedly shows that SPGISpeech model results in better WER, likely because of the larger amount of training data or more diverse accent representation in this corpus compared to LibriSpeech. Surprisingly, pre-trained Russian Open STT model yields even better results than SPGISpeech model, even if the amount of the training data for them is quite similar (about 5000 hours). Since Swahili language is not closely related to English or Russian, we attribute better results of Russian Open STT model either to the larger amount of acoustic conditions and speaking styles in Russian Open STT corpus, or to more similar output vocabulary in the model: both Russian and Swahili models use 100 subword units, while English models use 5000 units. Validation accuracy of wav2vec 2.0 models does not look promising in our experiments

[2] Using http://www.bantu-languages.com/en/tools/swahili_numbers.html

[3] https://github.com/snakers4/open_stt

and we do not include their decoding results to the table. Freezing the first Encoder layer of Russian Open STT model during training on Swahili data gives us consistent improvement on both testing datasets, but freezing more layers does not appear to be beneficial. Interestingly enough, external LM also improves results on both Coastal and Congolese Swahili, however the best LM weights differ between languages, and we conclude to keep them separate in the final system.

#	System	swa	swc	Avg.
1.	CTC weight 0.3	25.9	26.5	26.2
2.	CTC weight 0.4	25.8	24.4	25.1
3.	CTC weight 0.5	25.2	25.0	**25.1**
4.	CTC weight 0.6	25.4	25.0	25.2
5.	CTC weight 0.7	26.4	24.9	25.7
6.	#3, pre-trained LibriSpeech	22.4	25.4	23.9
7.	#3, pre-trained SPGISpeech	20.8	22.9	21.9
8.	#3, pre-trained Russian Open STT	21.4	20.8	**21.1**
9.	#8, freeze Encoder layers #1–4	20.3	21.1	20.7
10.	#8, freeze Encoder layers #1–2	21.9	21.4	21.7
11.	#8, freeze Encoder layer #1	17.8	19.7	**18.8**
12.	#11, average 9 checkpoints	17.7	19.7	18.7
13.	#11, average 8 checkpoints	17.7	19.5	**18.6**
14.	#11, average 7 checkpoints	17.8	19.6	18.7
15.	#11, average 6 checkpoints	17.7	19.5	18.6
16.	#11, average 5 checkpoints	17.9	19.6	18.8
17.	#13, external LM weight 0.2	15.1	18.4	16.8
18.	#13, external LM weight 0.3	14.5	**18.3**	16.4
19.	#13, external LM weight 0.4	14.0	18.5	16.3
20.	#13, external LM weight 0.5	13.6	18.7	16.2
21.	#13, external LM weight 0.6	**13.5**	19.1	16.3
22.	#13, external LM weight 0.7	13.8	19.9	16.9

Table 2: ASR results (WER, %) on the shared task validation data. Bold numbers correspond to the selected configuration for the final system (the external LM weights are language-specific).

3 MT

3.1 Data

Table 3 summarizes the datasets used to train our MT systems. The training data comprises of the shared task training data, Gamayun kit[4] (English – Swahili and Congolese Swahili – French parallel text corpora) as well as multiple corpora from the OPUS collection (Tiedemann, 2012), namely: ELRC_2922 (Tiedemann, 2012), GNOME (Tiedemann, 2012), CCAligned and MultiCCAligned (El-Kishky et al., 2020), EUbookshop (Tiedemann, 2012), GlobalVoices (Tiedemann, 2012), JW300 for sw and swc source languages (Agić and Vulić, 2019), ParaCrawl and MultiParaCrawl[5], Tanzil (Tiedemann, 2012), TED2020 (Reimers and Gurevych, 2020), Ubuntu (Tiedemann, 2012), WikiMatrix (Schwenk et al., 2019) and wikimedia (Tiedemann, 2012). The validation data for each target language comprises of 434 randomly sampled utterances from the shared task training data. The testing data is the shared task validation data, that also has 434 sentences per target language.

Dataset	Words		Sentences	
	→eng	→fra	→eng	→fra
IWSLT'21	31,594	51,111	4,157	4,562
Gamayun	39,608	216,408	5,000	25,223
ELRC_2922	12,691	-	607	-
GNOME	170	170	40	40
CCAligned	18,038,994	-	2,044,993	-
MultiCCAligned	18,039,148	10,713,654	2,044,991	1,071,168
EUbookshop	228	223	17	16
GlobalVoices	576,222	347,671	32,307	19,455
JW300 sw	15,811,865	15,763,811	964,549	931,112
JW300 swc	9,108,342	9,094,008	575,154	558,602
ParaCrawl	3,207,700	-	132,517	-
MultiParaCrawl	-	996,664	-	50,954
Tanzil	1,734,247	117,975	138,253	10,258
TED2020	136,162	134,601	9,745	9,606
Ubuntu	2,655	189	986	53
WikiMatrix	923,898	271,673	51,387	19,909
wikimedia	66,704	1,431	771	13
Total	41,910,113	30,003,158	3,406,772	2,159,007

Table 3: Datasets used to train the MT systems and their sizes in numbers of words (source language) and sentences. Total numbers are lower due to the deduplication.

3.2 Model

For the text-to-text neural machine translation (NMT) system we use a Transformer big model (Vaswani et al., 2017) using the fairseq implementation (Ott et al., 2019). We train three versions of the translation model.

First we train a vanilla NMT (`vanillaNMT`) system using only the data from the parallel training dataset. For preprocessing we use the SentencePiece implementation (Kudo and Richardson, 2018) of BPEs (Sennrich et al., 2016). For our second experiment for the NMT system (`preprocNMT`), we apply the same written to spoken language conversion as used for the ASR transcriptions (section §2.1) to the source text S and obtain ASR-like text S_t. S_t is then segmented using a BPE model and used as input for our NMT model. The last approach was using a multi-task framework to train the system (`multiNMT`), where all parameters of the translation model were shared. The main task of this model is to translate ASR output S_t to the target language T (task `asrS`), while our auxiliary task is to translate regular source Swahili S to the target language T (task `textS`). We base or multi-task approach on the idea of mul-

[4] https://gamayun.translatorswb.org/data/

[5] https://www.paracrawl.eu/

Model	Input	swa→eng		swc→fra		swc→eng	
		BLEU	chrF	BLEU	chrF	BLEU	chrF
vanillaNMT	textS	25.72	53.47	17.70	44.80	10.55	38.07
	asrS	14.26	47.74	10.57	40.99	4.71	34.70
	ASR #20	13.21	46.11	10.67	40.53	4.67	33.94
	ASR #1	11.50	43.34	9.52	38.32	4.24	32.45
preprocNMT	textS	11.01	41.33	13.54	41.00	4.49	31.91
	asrS	16.00	45.86	14.09	42.05	7.10	34.35
	ASR #20	14.54	44.17	13.23	41.00	6.62	33.63
	ASR #1	12.45	40.95	11.21	38.08	5.47	31.63
multiNMT	textS	25.69	53.27	18.20	44.66	10.56	38.29
	asrS	20.07	50.31	14.69	43.07	8.73	36.72
	ASR #20	17.91	48.39	13.29	41.58	7.94	35.47
	ASR #1	15.81	45.31	11.97	39.09	7.03	33.78

Table 4: MT results on the shared task validation data. WER values on swa/swc validation data are 13.6/18.7% for ASR #20 and 25.9/26.5% for ASR #1.

tilingual NMT introduced by Johnson et al. (2017), using a special token at the beginning of each sentence belonging to a certain task, as we can see in the next example:

<asrS> sara je haujui tena thamani ya kikombe hiki → Tu ne connais donc pas, Sarah, la valeur de cette coupe ?
<textS> Sara, je! Haujui tena, thamani ya kikombe hiki? → Tu ne connais donc pas, Sarah, la valeur de cette coupe ?

Then, our multi-task training objective is to maximize the joint log-likelihood of the auxiliary task textS and the primary task asrS.

Hyperparameters For word segmentation we use BPEs (Sennrich et al., 2016) with separate dictionaries for the encoder and the encoder, using the SentencePiece implementation (Kudo and Richardson, 2018). Both vocabularies have a size of 8000 tokens. Our model has 6 layers, 4 attention heads and embedding size of 512 for the encoder and the decoder. To optimize our model we use Adam (Kingma and Ba, 2014) with a learning rate of 0.001. Training was performed on 40 epochs with early stopping and a warm-up phase of 4000 updates. We also use a dropout (Srivastava et al., 2014) of 0.4, and an attention dropout of 0.2. For decoding we use Beam Search, with a size of 5.

3.3 Results

Table 4 shows the results of our MT system in combination with different inputs. We trained three models using the techniques described in section §3.2 (vanillaNMT, preprocNMT, and multiNMT). Then we used the official validation set as input (textS), and also applied asrS preprocessing. We used both inputs to test the performance of all models with different inputs. As expected, the vanillaNMT systems performs well with textS input (i.e 25.72 BLEU for swa→eng), but drops when using asrS. This pattern was later confirmed when using real ASR output (ASR #20 and ASR #1). We noticed, that training our model with asrS, instead of using textS improves slightly the results (i.e 16.00 BLEU with preprocNMT compared with 14.26 on vanillaNMT for swa→eng). But when we use multiNMT the performance strongly increase to 20.07 for swa→eng. This pattern also can be seen when using real ASR output (ASR #20 and ASR #1), and across all language pairs. We hypothesize that the multi-task framework helps the model to be more robust to different input formats, and allows it to generalize more the language internals.

4 End-to-End ST

4.1 Data

End-to-end ST is fine-tuned on the same speech recordings, as ASR data, but with transcriptions in English or in French. English and French transcriptions are obtained either from the datasets released with the shared task, or by running our MT system on Swahili transcriptions. External LMs for English and French outputs are trained on 10M sentences of the corresponding language from the OSCAR corpus (Ortiz Suárez et al., 2020).

4.2 Model

The end-to-end ST system comprises of the Encoder part of our ASR system and the whole MT system with removed input token embedding layer. All layers are frozen during the fine-tuning except of the top four layers of ASR Encoder and bottom three layers of MT Encoder. SpecAugment and gradient accumulation are disabled during the fine-tuning. Compared to the ASR system, end-to-end ST system has larger dictionary, what leads to shorter output sequences and allows us to increase the batch size to 60M bins. The rest of hyperparameters are the same as in the ASR system. We evaluate ST model separately and also with external LM that is set up as described in the ASR section.

4.3 Results

It can be seen from Table 5 that the end-to-end ST systems do not yet match the cascaded systems in translation quality in low resource settings. External LMs, however, slightly improve the results for both target languages.

Setting	`swa→eng`		`swc→fra`		`swc→eng`	
	BLEU	chrF	BLEU	chrF	BLEU	chrF
No LM	7.81	30.83	2.94	22.98	3.59	23.50
LM weight 0.4	8.82	31.26	3.73	22.07	4.06	23.89
LM weight 0.6	9.11	31.45	3.58	20.57	4.17	23.62

Table 5: End-to-end ST results on the shared task validation data.

5 Final systems

Table 6 shows validation scores of our final systems, as well as their evaluation scores provided by the organizers of the shared task. Our primary (cascaded) system here uses increased beam sizes: 30 for the ASR, 10 for the English MT and 25 for the French MT. `swc/swa` WERs of the final ASR systems are 12.5/17.6% on the validation sets. We did not observe improvement from the increased beam size on the contrastive systems and leave it at 2. It should be noted that the contrastive system is evaluated on incomplete output[6] for the `swc→fra` pair because of the technical issue on our side. We observe a large gap between the validation and evaluation scores for Coastal Swahili source language, what might indicate some sort of bias towards the validation set in our ASR or MT, or both. It is unclear why it does not happen for Congolese Swahili source language, because we optimized all our systems for the best performance on the validation sets for both source languages.

6 Conclusion

This paper described the IMS submission to the IWSLT 2021 Low-Resource Shared Task on Coastal and Congolese Swahili to English and French, explaining our intermediate ideas and results. Our system is ranked as the best for Congolese Swahili to French and English, and the second for Coastal Swahili to English. In spite of the simplicity of our cascade system, we show that the improving of ASR system with pre-trained models and afterward the tuning of MT system to optimize its fit to the ASR output achieves good results, even in challenging low resource settings. Additionally, we tried an end-to-end ST system with a lower performance. However, we learned that there is still room for improvement, and in future work we plan to investigate this research direction.

System	Set	`swa→eng`	`swc→fra`	`swc→eng`
Primary	Val.	18.3	13.7	7.9
(cascaded)	Eval.	14.9	13.5	7.7
Contrastive	Val.	9.1	3.7	4.0
(end-to-end)	Eval.	6.7	2.7	3.9

Table 6: Results (BLEU) of the primary and contrastive systems on the validation and evaluation data of the shared task.

7 Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC. This product contains or makes use of IARPA data, IARPA Babel Swahili Language Pack, IARPA-babel202b-v.1.0d (LDC2017S05). Manuel Mager received financial support by DAAD Doctoral Research Grant for this work.

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210.

Jess Andresen, Aric Bills, Thomas Conners, Eyal Dubinski, Jonathan G. Fiscus, Mary Harper, Kirill Kozlov, Nicolas Malyska, Jennifer Melot, Michelle Morrison, Josh Phillips, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Jamie Wong. 2017. IARPA Babel Swahili Language Pack IARPA-babel202b-v1.0d LDC2017S05.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian

[6] 406 of 2124 hypothesis are empty.

Herold. 2020. Start-before-end and end-to-end: Neural speech translation by apptek and rwth aachen university. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 44–54.

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

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

Hadrien Gelas, Laurent Besacier, and François Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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. *Proc. Interspeech 2020*, pages 5036–5040.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2020. Recent Developments on ESPnet Toolkit Boosted by Conformer. *arXiv preprint arXiv:2010.13956*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Naoyuki Kamo. 2021. ESPnet2 LibriSpeech recipe.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

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

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019*, pages 2613–2617.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

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*, 30:5998–6008.

Changhan Wang, Morgane Rivière, 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*.