NVIDIA NeMo Offline Speech Translation Systems for IWSLT 2023

Oleksii Hrinchuk[*1], **Vladimir Bataev**[1,2], **Evelina Bakhturina**[1], **Boris Ginsburg**[1]
[1]NVIDIA, Santa Clara, CA [2]University of London, London, UK

Abstract

This paper provides an overview of NVIDIA NeMo's speech translation systems for the IWSLT 2023 Offline Speech Translation Task. This year, we focused on end-to-end system which capitalizes on pre-trained models and synthetic data to mitigate the problem of direct speech translation data scarcity. When trained on IWSLT 2022 constrained data, our best En→De end-to-end model achieves the average score of 31 BLEU on 7 test sets from IWSLT 2010-2020 which improves over our last year cascade (28.4) and end-to-end (25.7) submissions. When trained on IWSLT 2023 constrained data, the average score drops to 29.5 BLEU.

1 Introduction

We participate in the IWSLT 2023 Offline Speech Translation Task (Agarwal et al., 2023) for English→German, English→Chinese, and English→Japanese. This year, we focus on an end-to-end model, which directly translates English audio into text in other languages.

In contrast to automatic speech recognition (ASR) and text-to-text neural machine translation (NMT), the data for direct speech translation (ST) is scarce and expensive. Thus, to train a high-quality end-to-end ST model, we heavily rely on a number of auxiliary models for which the amount of available data is enough. Specifically, we train the following models:

- ASR model with FastConformer-RNNT (Rekesh et al., 2023) architecture trained on all allowed data.
- NMT model with Transformer encoder-decoder architecture trained on all allowed bitext and in-domain fine-tuned on TED talks.
- Text-to-speech (TTS) model with FastPitch (Łańcucki, 2021) architecture trained on the English transcripts of TED talks.
- Supervised Hybrid Audio Segmentation (SHAS) model (Tsiamas et al., 2022) trained on TED talks.

Our constrained end-to-end ST model consists of a FastConformer encoder and a Transformer decoder. We initialize the encoder with the corresponding component from ASR and train our ST model on a mix of speech-to-text and text-to-text data. We replace all ground truth translations (wherever available) with synthetic ones generated with the NMT model and voice the English portion of parallel text corpora with TTS.

Our systems will be open-sourced as part of NVIDIA NeMo[1] framework (Kuchaiev et al., 2019).

2 Data

In this section, we describe the datasets used for training (Table 1). For evaluation, we used the development sets of Must-C v2 (Cattoni et al., 2021), as well as the test sets from past IWSLT competitions. We noticed that development data had a large overlap with training data, mostly because of the usage of the same TED talks in different datasets. Thus, we discarded all samples with overlapping transcripts and talk ids.

TED talks In the list of allowed data, there are several datasets comprised of TED talks, namely Must-C v1-v3, ST-TED (Jan et al., 2018), and TED-LIUM v3 (Hernandez et al., 2018) which have significant data overlap. After combining them together and doing deduplication, we ended up with the dataset of 370K unique samples (611 hours of English audio) we used for in-domain fine-tuning of various models. Further in the text, we refer to

[*]Correspondence to: ohrinchuk@nvidia.com

[1]https://github.com/NVIDIA/NeMo

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 442–448
July 13-14, 2023 ©2023 Association for Computational Linguistics

Table 1: Statistics of different datasets used for training our models in a `constrained` regime.

Model	**Segments (millions)**	**Time (hours)**
ASR	2.7	4800
NMT En→De	11	–
NMT En→Zh	7.5	–
NMT En→Ja	21	–
TTS	0.37	611

this dataset and its subsets with available translations to De/Zh/Ja as **TED talks**. See Table 2 for the detailed statistics of this dataset.

ASR For training our ASR model, we used LibriSpeech (Panayotov et al., 2015), Mozilla Common Voice v11.0 (Ardila et al., 2019), TED-LIUM v3 (Hernandez et al., 2018), VoxPopuli v2 (Wang et al., 2021), all available speech-to-English data from Must-C v1-v3 (Cattoni et al., 2021) En-De/Zh/Ja datasets, ST-TED (Jan et al., 2018), and Europarl-ST (Iranzo-Sánchez et al., 2020).

We converted all audio data to mono-channel 16kHz wav format. Of all the datasets allowed under the constrained submission, LibriSpeech and TED-LIUM v3 were the only datasets that provided transcripts with neither punctuation nor capitalization (P&C). For LibriSpeech, we managed to restore P&C from the dataset metadata available at their website[2]. For TED-LIUM v3, we applied P&C restoration model trained on the English portion of allowed bitext. Finally, we discarded all samples shorter than 0.2s and longer than 22s and all samples with transcripts present in the evaluation dataset. As a result, our training dataset contained 2.7M audio segments with a total duration of 4.8k hours.

MT For training our NMT models, we used all available bitext allowed for IWSLT 2023 constrained submission. After training, we additionally fine-tuned our models on bitexts from TED talks for each language.

We applied `langid` and `bicleaner` filtering following Subramanian et al. (2021) and discarded all sentences longer than 128 tokens and sentences with the length ratio between source and target exceeding 3. We also applied Moses tokenization for En/De, jieba tokenization for Zh, and ja-mecab tokenization for Ja.

Table 2: Statistics of TED talks dataset.

Model	**Segments (thousands)**	**Time (hours)**
En audio → En text	370	611
En audio → De text	280	459
En audio → Zh text	350	580
En audio → Ja text	321	528

TTS For training our TTS model, we used TED talks with English transcripts. The combination of Must-C v1-v3 and ST-TED contained 3696 speakers, however, some of them were not unique. Capitalizing on the huge overlap with TED-LIUM v3 and the speaker names from there, we managed to attribute several talks to a single speaker reducing the number of unique speakers to 3361. We also removed capitalization from English transcripts in TED talks.

ST For training our end-to-end ST models, we used the combination of 1) ASR data with the ground truth transcripts replaced by synthetic translations; 2) NMT data with TTS-generated English audios on source side (Table 1).

3 System

In this section, we describe the essential components of our end-to-end submission.

ASR We trained 17-layer large conformer-transducer (Gulati et al., 2020) with FastConformer (Rekesh et al., 2023) encoder and RNN-T loss and decoder (Graves, 2012). The prediction network consisted of a single layer of LSTM (Hochreiter and Schmidhuber, 1997), and the joint network is an MLP. All the hidden sizes in the decoder were set to 640. Unigram SentencePiece (Kudo and Richardson, 2018) with 1024 tokens was used for tokenization.

The ASR models were trained for 45 epochs, starting with a checkpoint pre-trained on LibriSpeech. We used AdamW (Loshchilov and Hutter, 2017) optimizer and Noam Annealing (Vaswani et al., 2017) with 10K warmup steps and a maximum learning rate of 1.15. Weight decay of 0.001 on all parameters was used for regularization. The

[2] https://www.openslr.org/12

effective batch size was set to 1200, and we could fit larger batch sizes via batch splitting for the RNN-T loss. Time-Adaptive SpecAugment (Park et al., 2020) with 2 freq masks ($F = 27$) and 10 time masks ($T = 5\%$) was used as the augmentation scheme. We also used dropout of 0.1 for both the attention scores and intermediate activations.

NMT We trained our NMT models (Transformer, 12×6 layers, $d_{\text{model}} = 1024$, $d_{\text{inner}} = 4096$, $n_{\text{heads}} = 16$) with Adam optimizer (Kingma and Ba, 2014) and inverse square root annealing (Vaswani et al., 2017) with 7.5K warmup steps and a maximum learning rate of 10^{-3}. The models were trained for a maximum of 75K steps with a dropout of 0.1 on intermediate activations and label smoothing with $\alpha = 0.1$. Our En→De models used joint BPE vocabulary of 16384 tokens and En→Zh/Ja used separate vocabularies with the same number of tokens per language.

After training, we did checkpoint averaging and fine-tuned all our base NMT models on TED talks for 3 epochs with an initial learning rate of 2×10^{-5}, inverse square root annealing, and a warmup of 10% steps. Finally, we ensembled 2 models trained with different initializations for each language direction.

TTS Our TTS model was multi-speaker FastPitch (Łańcucki, 2021) text-to-mel-spectrogram generator. Training vocoder was not necessary for our setup as the parameters of spectrograms matched ones for ST models following the approach described in (Bataev et al., 2023). TTS-generated spectrograms were fed directly into the FastConformer encoder when training the ST model. Our TTS model was trained for 200 epochs on TED talks with restored speakers from TED-LIUM v3 (Hernandez et al., 2018).

Segmentation We used Supervised Hybrid Audio Segmentation (SHAS) approach following Tsiamas et al. (2022). As using speech representation pre-trained wav2vec 2.0 (Baevski et al., 2020) goes beyond the scope of `constrained` submission, we replaced it with Conformer ASR encoder, pre-trained on LibriSpeech.

ST Our end-to-end model consisted of FastConformer encoder followed by Transformer trained on pairs of English audio and transcripts in other languages (17-layer FastConformer encoder, 6×6 Transformer, both with $d_{\text{model}} = 512$, $d_{\text{inner}} = 2048$, $n_{\text{heads}} = 8$). We used the vocabulary of 16384 YouTokenToMe[3] byte-pair-encodings, trained jointly for En→De and separately for En→Zh/Ja. All models were trained for 30k steps with ASR-initialized encoder and randomly initialized decoder.

Table 3: Word error rate (WER) of the English ASR model evaluated on TED talks from Must-C v2 and past test sets from IWSLT. All predictions and ground truths transcripts were normalized for WER computation.

Model	**tst-COM**		**IWSLT.tst**		
	De	Zh/Ja	2018	2019	2020
norm	5.9	5.8	9.8	5.6	8.0
punct	5.7	5.4	9.4	4.9	7.0
punct+capit	5.7	5.5	9.5	5.7	8.5

To speed up training and improve GPU utilization, we bucketed our ASR and NMT datasets on sequence length so each batch contained a similar number of tokens. On each iteration, we pick one batch from ASR and one batch which resulted in approximately 3:2 ratio between segments from ASR and NMT for En→De. TTS mel spectrograms were generated on-the-fly for a randomly selected speaker for each sample.

After pretraining on the ASR task, we fused BatchNorm in FastConformer layers as proposed in (Bataev et al., 2023) to avoid a mismatch between statistics for natural and generated mel spectrograms. The batch normalization layer was replaced with a trainable projection initialized from the original parameters. We observed meaningful improvements when using such an approach compared to retaining the original batch normalization.

4 Experiments

4.1 Results

ASR Table 3 shows word error rate (WER) of our ASR models on different evaluation datasets. We trained 3 models which differed by the format of transcripts: normalized (`norm`), with punctuation only (`punct`), with punctuation and capitalization (`punct+capit`).

All models exhibited similar results, with `punct` being slightly better on all evaluation datasets. However, in our further experiments of training end-to-end ST with an ASR-initialized en-

[3] https://github.com/VKCOM/YouTokenToMe

Table 4: En→De BLEU scores calculated on IWSLT test sets from different years by using automatic re-segmentation of the hypothesis based on the reference translation by `mwerSegmenter` implemented in SLTev (Ansari et al., 2021). Avg Δ computes the improvement over the cascade baseline averaged over 7 test sets.

Model description	**2010**	**2013**	**2014**	**2015**	**2018**	**2019**	**2020**	**Avg**
				Text-to-text NMT models				
Transformer 12 × 6 `constrained`	32.9	36.7	32.7	34.2	30.5	29.4	33.0	32.8
+ checkpoint averaging	33.1	37.4	32.8	35.1	30.3	29.8	33.5	33.1
+ TED talks fine-tuning	34.5	39.1	34.1	35.3	30.8	30.3	33.8	34.0
+ x2 ensembling	35.2	40.2	34.9	36.0	32.5	31.6	35.4	35.1
NeMo IWSLT'22 NMT model	35.7	41.2	36.2	38.1	34.7	31.7	35.0	36.1
				End-to-end ST models				
Conformer (17) + Transformer (6 × 6)	29.8	33.8	30.2	27.1	26.2	26.8	29.1	29.0
+ better WebRTC VAD parameters	31.2	35.4	31.8	28.6	27.3	27.6	29.7	30.2
+ SHAS segmentation	32.1	36.1	32.6	29.0	28.4	27.9	30.9	31.0
NeMo IWSLT 2023 `constrained`	31.0	34.9	30.7	28.6	27.4	27.7	30.3	29.5
NeMo IWSLT 2022 (end-to-end)	24.5	30.0	25.2	25.3	24.9	24.1	26.2	25.7
NeMo IWSLT 2022 (cascade)	26.6	32.2	26.8	28.3	28.1	27.3	29.7	28.4
KIT IWSLT 2022	–	–	–	27.9	–	27.6	30.0	–
USTC-NELSLIP IWSLT 2022	–	–	–	–	29.9	28.2	30.6	–
YiTrans IWSLT 2022	–	–	–	–	–	31.6	34.1	–

coder, we did not notice a significant difference in the corresponding BLEU scores.

ST En→De Table 4 shows the performance of our baseline En→De system and its ablations on 7 different IWSLT test sets over the years. All ablation experiments used the last year's constrained setup that included more NMT data from WMT to be comparable with the last year submissions. The systems we submit were retrained on the allowed data to comply with `constrained` restrictions.

We improve the average BLEU score by 5.3 over our last year end-to-end submission. We believe that such gain is attributed to several factors, most importantly, switching to synthetic transcripts, including TTS-generated data, and a better segmentation model. On some of the evaluation datasets, we approached the BLEU scores of top contestants from last year.

Retraining our model in accordance with this year `constrained` setup resulted in the average degradation of 1.5 BLEU. Most of this performance drop was attributed to worse NMT models trained on limited amount of data which did not include large bitexts from WMT.

ST En→Zh/Ja To train English-Chinese and English-Japanese ST systems, we followed a similar recipe to the English-German system. Specifically, we re-trained NMT components and used them to generate synthetic translations of audio segments. With other auxiliary models intact, we replaced bitexts used for TTS augmentations and trained En→Zh (Table 5) and En→Ja (Table 6) ST end-to-end models in a `constrained` setup.

The only difference in our submission was that the English-Chinese model used `punct+capit` ASR, while the English-Japanese model used `norm` ASR. This choice was based on a slightly higher (less than 0.5) BLEU score on Must-C v2 dev dataset.

4.2 Discarded alternatives

When designing our submission, we explored a number of alternatives that did not lead to a clear improvement in preliminary experiments and, thus, were not included in the final submission.

ASR We tried to replace BatchNorm with LayerNorm in the FastConformer backbone to mitigate the statistics mismatch between natural and TTS-generated mel-spectrograms. The resulting model

Table 5: En→Zh BLEU scores calculated on Must-C `dev` and `tst-COMMON` with official segmentation.

Model description	**dev**	**tst-COM**
Text-to-text NMT models		
Transformer 12 × 6	22.9	26.4
+ ckpt avg	23.0	26.4
+ TED talks fine-tuning	24.7	28.0
+ x2 ensembling	25.5	28.9
End-to-end ST models		
NeMo IWSLT 2023	23.9	27.5
USTC-NELSLIP IWSLT'22	–	28.7
YiTrans IWSLT'22	–	29.3

Table 6: En→Ja BLEU scores calculated on Must-C `dev` and `tst-COMMON` with official segmentation.

Model description	**dev**	**tst-COM**
Text-to-text NMT models		
Transformer 12 × 6	12.8	15.5
+ ckpt avg	13.3	16.2
+ TED talks fine-tuning	14.7	18.5
+ x2 ensembling	15.0	19.2
End-to-end ST models		
NeMo IWSLT 2023	14.5	18.3
USTC-NELSLIP IWSLT'22	–	18.2
YiTrans IWSLT'22	–	19.1

required more epochs to converge and resulted in slightly higher WER.

NMT We experimented with larger models of up to 12 × 8 layers, larger vocabularies of up to 32k tokens, and label smoothing of up to 0.2 but did not notice any improvements to BLEU scores. We also saw diminishing returns when using more than 2 models in the ensemble. Thus, we decided to stick to the ensemble of two 12 × 6 models with 16k vocab to speed up synthetic data generation.

TTS While debugging the code, we noticed that TTS model generating mel-spectrograms used the same single speaker and had dropout enabled. Surprisingly, it did not lead to performance degradation. We hypothesize that this was caused by using well converged pre-trained ASR encoder, which was not altered significantly by the low-quality signal. We also experimented with improving generated spectrograms with GAN enhancer following Bataev et al. (2023), which led to similar results at the cost of significant computation overhead.

Segmentation We experimented with voice activity detection implemented in `WebRTC`[4] toolkit, however, the BLEU scores on IWSLT test sets were lower even after extensive hyperparameter search.

ST Given the effectiveness of ensembling in last year's competition, we evaluated the performance of an ensemble of up to 3 models with different ASR encoder initializations. Unlike NMT, we did not observe any improvement in using the best model from the ensemble.

We experimented with using RNN-T instead of the Transformer decoder. Despite its remarkable performance in ASR, RNN-T converged much slower and underperformed our Transformer decoder by more than 2 BLEU in our ST model.

5 Conclusion

We present NVIDIA NeMo group's offline speech translation systems for En→De, En→Zh, and En→Ja IWSLT 2023 Tasks.

Our *primary* end-to-end models that translate English speech directly into German, Chinese, and Japanese texts, consist of FastConformer encoder and Transformer decoder. To alleviate the problem of direct ST data scarcity, we capitalized on a number of auxiliary ASR, TTS, and NMT models, and their ability to generate hiqh-quality audio and translations. The resulting models achieve competitive performance without using any amount of direct ST data.

Although we participated in `constrained` scenario, our pipeline can be easily scaled to arbitrarily large amounts of ASR and NMT data.

Acknowledgments

The authors would like to thank Somshubra Majumdar for many useful discussions over the course of this project and Nithin Koluguri for help with training ASR models.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda

[4] https://github.com/wiseman/py-webrtcvad

Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

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*, 33:12449–12460.

Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator. *ArXiv*, abs/2302.14036.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for speech recognition. In *Proceedings of Interspeech*, pages 5036–5040.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria 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. IEEE.

Niehues Jan, Roldano Cattoni, Stüker Sebastian, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *Proceedings of IWSLT*, pages 2–6.

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

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

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*.

Adrian Łańcucki. 2021. FastPitch: Parallel text-to-speech with pitch prediction. In *ICASSP*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210. IEEE.

Daniel S Park, Yu Zhang, Chung-Cheng Chiu, Youzheng Chen, Bo Li, William Chan, Quoc V Le, and Yonghui Wu. 2020. Specaugment on large scale datasets. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6879–6883. IEEE.

Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *arXiv preprint arXiv:2305.05084*.

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. Nvidia nemo neural machine translation systems for english-german and english-russian news and biomedical tasks at wmt21. *arXiv preprint arXiv:2111.08634*.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

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*.