FAIRSEQ S2T: Fast Speech-to-Text Modeling with FAIRSEQ

Changhan Wang[1], Yun Tang[1], Xutai Ma[1,2], Anne Wu[1], Dmytro Okhonko[1], Juan Pino[1]

[1]Facebook AI
[2]Johns Hopkins University

xutai_ma@jhu.edu
{changhan,yuntang,xutaima,annewu,oxo,juancarabina}@fb.com

Abstract

We introduce FAIRSEQ S2T, a FAIRSEQ (Ott et al., 2019) extension for speech-to-text (S2T) modeling tasks such as end-to-end speech recognition and speech-to-text translation. It follows FAIRSEQ's careful design for scalability and extensibility. We provide end-to-end workflows from data pre-processing, model training to offline (online) inference. We implement state-of-the-art RNN-based as well as Transformer-based models and open-source detailed training recipes. FAIRSEQ's machine translation models and language models can be seamlessly integrated into S2T workflows for multi-task learning or transfer learning. FAIRSEQ S2T documentation and examples are available at https://github.com/pytorch/fairseq/tree/master/examples/speech_to_text.

1 Introduction

End-to-end sequence-to-sequence (S2S) modeling has witnessed rapidly increased applications in speech-to-text (S2T) tasks. It achieves state-of-the-art performance on automatic speech recognition (ASR) (Park et al., 2019; Synnaeve et al., 2019) and leads to the recent resurgence of speech-to-text translation (ST) research (Duong et al., 2016; Bérard et al., 2016). ASR and ST are closely related. There are recent attempts to combine the two tasks under the same S2S model architecture via multi-task learning (Anastasopoulos and Chiang, 2018; Liu et al., 2020). They also benefit from each other via transfer learning (Bansal et al., 2019; Wang et al., 2020b) and are able to leverage additional supervision from machine translation (MT) and language modeling (LM). When supervised data is not abundant, self-supervised pretraining (Schneider et al., 2019; Wu et al., 2020) and semi-supervised training (Kahn et al., 2020; Pino et al., 2020) lowers the requirements on supervision and improves model performance.

The increased connections among ASR, ST, MT and LM has called for all-in-one S2S modeling toolkits, and the use of large-scale unlabeled speech data sets the scalability requirements. In this paper, we introduce FAIRSEQ S2T, a FAIRSEQ (Ott et al., 2019) extension for S2T tasks such as end-to-end ASR and ST. It follows FAIRSEQ's careful design for scalability and extensibility. We provide end-to-end workflows from data pre-processing, model training to offline (online) inference. We implement state-of-the-art RNN-based (Chan et al., 2016; Bérard et al., 2018) and Transformer-based (Vaswani et al., 2017; Mohamed et al., 2019) models and open-source detailed training recipes. FAIRSEQ's MT models and LMs can be seamlessly integrated into S2T workflows for multi-task learning or transfer learning. To facilitate model evaluation, we add a collection of scorers as well as VizSeq (Wang et al., 2019) integration for visualized error analysis. FAIRSEQ S2T documentation and examples are available at https://github.com/pytorch/fairseq/tree/master/examples/speech_to_text.

With counterpart toolkits such as ESPNet (Inaguma et al., 2020) and Lingvo (Shen et al., 2019), FAIRSEQ S2T pursues the best integration, scalability and reproducibility. A detailed comparison of FAIRSEQ S2T with its counterparts can be found in Table 1.

2 Features

Fairseq Models FAIRSEQ provides a collection of MT models (Ng et al., 2019; Lewis et al., 2020) and LMs (Liu et al., 2019; Conneau et al., 2020) that demonstrate state-of-the-art performance on standard benchmarks. They are open-sourced with pre-trained models. FAIRSEQ also supports other tasks such as text summarization, story generation and self-supervised speech pre-training.

	ASR	LM	MT	Non-Autoreg. MT	Offline ST	Online ST	Speech Pre-training	Multi-node training	Pre-trained models
ESPNet-ST	✓	✓	✓		✓			✓[†]	✓
Lingvo	✓	✓	✓		✓[‡]			✓	
OpenSeq2seq[1]	✓	✓	✓					✓	✓
RETURNN[2]	✓	✓	✓		✓			✓	✓
SLT.KIT[3]	✓		✓		✓				✓
Tensor2Tensor[4]	✓	✓	✓					✓	✓
OpenNMT[5]	✓	✓	✓					✓	✓
Kaldi[6]	✓	✓							✓
Wav2letter++[7]	✓	✓							✓
fairseq S2T	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of FAIRSEQ S2T with counterpart toolkits (as of July 2020). [†] Only available in version 2 (under development). [‡] Not publicly available. [1] Kuchaiev et al. (2018). [2] Zeyer et al. (2018). [3] Zenkel et al. (2018). [4] Vaswani et al. (2018). [5] Klein et al. (2017). [6] Povey et al. (2011). [7] Pratap et al. (2018).

S2T extension FAIRSEQ S2T adds attention-based RNN models (Chan et al., 2016; Bérard et al., 2018) as well as the latest Transformer models (Vaswani et al., 2017; Mohamed et al., 2019) for ASR and ST. It also supports CTC criterion (Graves et al., 2006) for ASR. For the simultaneous ST setting, it includes online models with widely used policies: monotonic attention (Raffel et al., 2017), wait-k (Ma et al., 2019), monotonic infinite look-back attention (Arivazhagan et al., 2019b), and monotonic multihead attention (Ma et al., 2020b).

Data Pre-Processing FAIRSEQ S2T extracts Kaldi-compliant (Povey et al., 2011) speech features (e.g. log mel-filter banks) automatically from WAV/FLAC audio files via PyKaldi (Can et al., 2018) or torchaudio[1]. Speech features can also be pre-computed and stored in NumPy (Harris et al., 2020) format. Optionally, raw audio files or features files can be packed into ZIP archives to improve I/O performance or facilitate file management. For further pre-processing, FAIRSEQ S2T provides online speech data transforms, including CMVN (cepstral mean and variance normalization), speed perturbation (Ko et al., 2017) and SpecAugment (Park et al., 2019). It also has an open interface for user-defined transforms. For text data, FAIRSEQ S2T does online tokenization with a rich collection of tokenizers, including Moses[2], SentencePiece (Kudo and Richardson, 2018), subword-nmt[3], byte-level BPE (Wang et al., 2020a) and bytes (Li et al., 2019).

Data Configuration FAIRSEQ S2T gets raw audio (feature) paths and target texts from manifest files in TSV (tab-separated values) format, which is similar to Kaldi-style scp files. Online speech data transforms and other data-related settings (e.g. tokenizer type and vocabulary) are defined by a separate configuration file in YAML format.

Computation FAIRSEQ is implemented in Py-Torch (Paszke et al., 2019) and it provides efficient batching, mixed precision training (Micikevicius et al., 2018), multi-GPU as well as multi-machine training for computational efficiency on large-scale experiments.

Evaluation Metrics FAIRSEQ S2T provides common automatic metrics for ASR, ST and MT, including WER (word error rate), BLEU (Papineni et al., 2002) and chrF (Popović, 2015). It also integrates SIMULEVAL (Ma et al., 2020a) for simultaneous ST/MT metrics such as AL (average lagging) (Ma et al., 2019) and DAL (differentiable average Lagging) (Cherry and Foster, 2019).

Visualization FAIRSEQ supports Tensorboard[4] for monitoring holistic metrics during model training. It also has VizSeq (Wang et al., 2019) integration for sequence-level error analysis, where speech and target/predicted text data are visualized with alignments in Jupyter Notebook interface.

3 Experiments

We evaluate FAIRSEQ S2T models on English ASR benchmark—LibriSpeech (Panayotov et al., 2015), as well as multilingual ST benchmarks—MuST-C (Di Gangi et al., 2019a) and CoVoST 2 (Wang

---

[1]https://github.com/pytorch/audio
[2]https://github.com/moses-smt/mosesdecoder
[3]https://github.com/rsennrich/subword-nmt

[4]https://github.com/tensorflow/tensorboard

	De	Nl	Es	Fr	It	Pt	Ro	Ru
Transformer[1]	17.3	18.8	20.8	26.9	16.8	20.1	16.5	10.5
Transformer[2][†]	22.9	27.4	28.0	32.7	23.8	28.0	21.9	15.8
T-Sm	22.7	27.3	27.2	32.9	22.7	28.1	21.9	15.3
Multi. T-Md[*]	24.5	28.6	28.2	34.9	24.6	31.1	23.8	16.0
Offline	19.2	23.5	24.0	29.1	16.4	23.5	19.7	13.7
High Lat.[‡]	18.6 (6.8)	22.9 (6.9)	22.3 (6.8)	28.4 (6.7)	15.4 (6.8)	22.6 (6.9)	19.1 (6.7)	12.9 (6.9)
Mid Lat.[‡]	14.1 (5.4)	17.9 (5.4)	17.2 (5.5)	25.0 (5.3)	12.0 (5.5)	17.7 (5.8)	15.0 (5.6)	7.2 (5.8)
Low Lat.[‡]	8.2 (2.9)	12.3 (2.8)	13.0 (3.0)	21.1 (2.8)	6.7 (2.9)	13.3 (2.9)	12.1 (2.9)	4.9 (2.7)

Table 2: FAIRSEQ S2T models on MuST-C. Test BLEU reported (for online models, AL is shown in parentheses). [1] Di Gangi et al. (2019). [2] Inaguma et al. (2020). [†] Applied additional techniques: speed perturbation, pre-trained decoder from MT and auxiliary CTC loss for ASR pre-training. [‡] Online models using beam size of 1 (instead of 5). [*] Trained jointly on all 8 languages.

	Type	Config.	Params
B-Base	RNN[†]	512d, 3L enc./2L dec.	31M
B-Big		512d, 5L enc./3L dec.	52M
T-Sm	Transformer[‡]	256d, 12L enc./6L dec.	31M
T-Md		512d, 12L enc./6L dec.	72M
T-Lg		1024d, 12L enc./6L dec.	263M

Table 3: FAIRSEQ S2T models for benchmarking. For simplicity, we use the same (default) model hyper-parameters and learning rate schedule across all experiments. [†] Bérard et al. (2018). [‡] Vaswani et al. (2017).

	Dev		Test	
	Clean	Other	Clean	Other
LAS[†]	-	-	2.8	6.8
Transformer[‡]	2.5	6.7	2.9	7.0
B-Big	3.7	11.4	3.9	11.5
T-Sm	4.1	9.3	4.4	9.2
T-Md	3.5	8.1	3.7	8.1
T-Lg	3.3	7.7	3.5	7.8

Table 4: FAIRSEQ S2T models on LibriSpeech (using default hyper-parameters and LR schedule). Dev and test WER reported. [†] Park et al. (2019). [‡] Synnaeve et al. (2019).

et al., 2020c). The model architectures used in benchmarking can be found in Table 3.

3.1 Experimental Setup

For speech inputs, we extract 80-channel log mel-filter bank features (25ms window size and 10ms shift) with utterance-level CMVN applied. We remove training samples with more than 3,000 frames for GPU memory efficiency. To alleviate overfitting, we pre-train ST model encoders on English ASR and adopt SpecAugment (without time warping): LD policy on LibriSpeech models and LB policy on MuST-C and CoVoST 2 models. We av-

erage the last 10 checkpoints and use a beam size of 5 for decoding. For ASR, we use 10K unigram vocabulary (Kudo and Richardson, 2018) and report WER. For ST, we use character vocabulary for CoVoST 2 and 8K unigram vocabulary for MuST-C. We report case-sensitive detokenized BLEU using sacreBLEU (Post, 2018), except for Japanese and Chinese translations (no word segmentation) where we report character-level BLEU.

3.2 Speech Recognition (ASR)

LibriSpeech is a de-facto standard ASR benchmark that contains 1,000 hours of English speech from audiobooks. Table 4 shows the dev and test WER of our models on LibriSpeech clean and noisy sets. Two popular architectures, RNN-based model ("B-Big") and Transformer based models ("T-Sm", "T-Md" and "T-Lg"), are evaluated. We can see that both architectures are able to achieve competitive performance (WER) to the state-of-the-art ones (the upper section), while we use only default model hyper-parameters and learning rate schedule without any task-specific tuning.

3.3 Speech Translation (ST)

3.3.1 MuST-C

MuST-C contains up to around 500 hours of English speech from TED talks with translations in 8 European languages. Table 2 shows the test BLEU of our Transformer-based models ("T-Sm" and "Multi. T-Md") and RNN-based models ("B-Base") on all the MuST-C language directions. Compared with previous Transformer-based approaches (Di Gangi et al., 2019b; Inaguma et al., 2020), our bilingual models achieve comparative results to the state of the art without applying additional techniques such as speed perturbation and

	Fr	De	Es	Zh	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
X→En													
B-Base	23.2	15.7	20.2	4.4	2.2	2.7	1.4	1.2	1.5	0.2	1.1	1.0	1.7
+ SSL[*]	23.1	16.2	20.2	4.8	3.2	3.8	3.7	2.3	2.2	0.2	1.6	1.6	2.2
Multi. B-Big[†]	26.6	19.5	26.3	4.4	2.1	0.3	1.3	0.6	1.4	0.1	0.6	0.3	0.9
T-Sm	26.3	17.1	23.0	5.8	3.6	4.3	2.7	2.5	3.0	0.3	1.5	2.5	2.7
Multi. T-Md[†]	26.5	17.5	27.0	5.9	2.3	0.4	0.5	0.6	0.7	0.1	0.1	0.3	1.9
En→X													
B-Base	-	12.5	-	20.0	6.7	9.1	18.1	8.7	11.6	7.4	25.6	15.2	18.9
Multi. B-Big[‡]	-	12.6	-	22.2	7.3	8.0	18.3	8.9	11.4	7.3	28.2	16.0	19.3
T-Sm	-	16.3	-	25.4	10.0	12.1	21.8	13.0	16.0	10.9	29.6	20.4	23.9
Multi. T-Md[‡]	-	15.4	-	26.5	9.5	10.8	20.9	12.2	14.6	10.3	30.5	18.9	22.0

Table 5: FAIRSEQ S2T models on CoVoST 2. Test BLEU reported (character-level BLEU for Zh and Ja targets). [*] Replaced mel-filter bank features with wav2vec ones (Schneider et al., 2019; Wu et al., 2020). [†] Trained jointly on all 21 X-En directions with temperature-based (T=2) resampling (Arivazhagan et al., 2019a). [‡] Trained jointly on all 15 En-X directions.

pre-trained decoder from MT. Moreover, our multilingual model (trained on all 8 languages) outperforms all bilingual ones with large margins. Besides traditional offline models, we also provide simultaneous ST models: the lower section in Table 2 presents the online models with wait-k policy, which was the baseline system in the IWSLT 2020 shared task on simultaneous ST (Ansari et al., 2020). The results represent the best systems in high (AL > 6), medium ($6 \geq$ AL > 3) and low (AL $\leq$ 3) latency regimes, on which we can clearly see the trade-offs between model performance and prediction latency.

3.3.2 CoVoST 2

CoVoST 2 contains total 2,880 hours of read speech in 22 languages from the open-source community, with 21 X-En directions and 15 En-X directions. We evaluate our models bidirectionally on 13 languages of them, including low-resource X-En directions: Zh, Tr, Ar, Sv, Lv, Sl, Ta, Ja, Id and Cy. We observe from Table 5 that our Transformer-based models ("T-Sm" and "T-Md") outperforms RNN-based ones ("B-Base" and "B-Big") on all En-X and X-En directions. The performance gap tends to be larger when the training data is higher resource (En-X directions, Fr-En, De-En and Es-En). Our multilingual models perform reasonably well with a universal model for over 15 X-En or En-X directions. They even have significant improvements on some directions (e.g. at least 4 BLEU gain on Es-En). For low-resource directions, we also evaluate self-supervised speech features (Schneider et al.,

2019; Wu et al., 2020)[5] as an alternative to the traditional log mel-filter bank features ("+ SSL"). We find that self-supervised features bring consistent gains and transfer well across different languages (self-supervised model trained on English and feature extracted for non-English).

4 Conclusion

We introduce FAIRSEQ S2T, a FAIRSEQ extension for speech-to-text (S2T) modeling tasks such as speech recognition and speech translation. It includes end-to-end workflows and state-of-the-art models with scalablity and extensibility design. It seamlessly integrates FAIRSEQ's machine translation models and language models to improve S2T model performance. FAIRSEQ S2T documentation and examples are available at `https://github.com/pytorch/fairseq/tree/master/examples/speech_to_text`.

Acknowledgments

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

---

[5]From a wav2vec model pre-trained on LibriSpeech: `https://github.com/pytorch/fairseq/tree/master/examples/wav2vec`

Volume 1 (Long Papers), pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019a. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019b. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pretraining on high-resource speech recognition improves low-resource speech-to-text translation. 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 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Dogan Can, Victor R. Martinez, Pavlos Papadopoulos, and Shrikanth S. Narayanan. 2018. Pykaldi: A python wrapper for kaldi. In *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

M. A. Di Gangi, M. Negri, and M. Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592.

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

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 21–31, Dublin, Ireland. European Association for Machine Translation.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959.

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.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics: System Demonstrations, pages 302–311, Online. Association for Computational Linguistics.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. 2018. Openseq2seq: extensible toolkit for distributed and mixed precision training of sequence-to-sequence models. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 41–46.

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, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, and William Chan. 2019. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5621–5625. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8417–8424.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. Simuleval: An evaluation toolkit for simultaneous translation. *arXiv preprint arXiv:2007.16193*.

Xutai Ma, Juan Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020b. Monotonic multihead attention. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, Conference Track Proceedings*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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. *Interspeech 2019*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. *arXiv preprint arXiv:2006.02490*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2018. wav2letter++: The fastest open-source speech recognition system. *CoRR*, abs/1812.07625.

Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *International Conference on Machine Learning*, pages 2837–2846.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, et al. 2019. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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 *Advances in neural information processing systems*, pages 5998–6008.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020a. Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9154–9160.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. VizSeq: a visual analysis toolkit for text generation tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 253–258, Hong Kong, China. Association for Computational Linguistics.

Changhan Wang, Juan Pino, and Jiatao Gu. 2020b. Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation. *arXiv preprint arXiv:2006.05474*.

Changhan Wang, Anne Wu, and Juan Pino. 2020c. Covost 2 and massively multilingual speech-to-text translation. *arXiv e-prints*, pages arXiv–2007.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. *arXiv preprint arXiv:2006.12124*.

Thomas Zenkel, Matthias Sperber, Jan Niehues, Markus Müller, Ngoc-Quan Pham, Sebastian Stüker, and Alex Waibel. 2018. Open source toolkit for speech to text translation. *The Prague Bulletin of Mathematical Linguistics*, 111(1):125–135.

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Proceedings of ACL 2018, System Demonstrations*, pages 128–133, Melbourne, Australia. Association for Computational Linguistics.