MAKE MORE OF YOUR DATA: MINIMAL EFFORT DATA AUGMENTATION FOR AUTOMATIC SPEECH RECOGNITION AND TRANSLATION

Tsz Kin Lam[⋆] *Shigehiko Schamoni*[⋆,†] *Stefan Riezler*[⋆,†]

[⋆]Computational Linguistics & [†]IWR, Heidelberg University, Germany
{lam,schamoni,riezler}@cl.uni-heidelberg.de

ABSTRACT

Data augmentation is a technique to generate new training data based on existing data. We evaluate the simple and cost-effective method of concatenating the original data examples to build new training instances. Continued training with such augmented data is able to improve off-the-shelf Transformer and Conformer models that were optimized on the original data only. We demonstrate considerable improvements on the LibriSpeech-960h test sets (WER 2.83 and 6.87 for `test-clean` and `test-other`), which carry over to models combined with shallow fusion (WER 2.55 and 6.27). Our method of continued training also leads to improvements of up to 0.9 WER on the ASR part of CoVoST-2 for four non-English languages, and we observe that the gains are highly dependent on the size of the original training data. We compare different concatenation strategies and found that our method does not need speaker information to achieve its improvements. Finally, we demonstrate on two datasets that our methods also works for speech translation tasks.

Index Terms— automatic speech recognition, speech translation, data augmentation, on-the-fly, resource efficient

1. INTRODUCTION

Data augmentation (DA) is an active research field in many machine learning areas. It addresses the problem of creating large and informative datasets for data-hungry neural networks with automatic techniques. The standard technique for DA is extending or enhancing training data, such that the final model quality is improved either by simply increasing the total amount of training data, or by creating informative training instances that improve robustness. In practice, both aspects are interwoven and work together.

The situation in speech-to-text processing such as automatic speech recognition (ASR) is particularly difficult due to two main reasons: First, generating speech data sets is expensive, and second, training of ASR models is extremely resource intensive. Thus, it is desirable to make most of the data and of pre-trained models that are available.

We evaluate the applicability of one of simplest DA techniques, namely concatenating training instances of the original data to create new training instances, to speech-to-text processing. Our method does not need any additional data or resources, and comes with low computational effort that allows applying the augmentation procedure in-memory and on-the-fly. Our experiments show that already very strong models can be further improved with continued training using a concatenation based DA approach. We further evaluate different strategies for selecting data to concatenate, and find that these strategies can make a difference depending on the size and complexity of the data set. Furthermore, we show that it is important to combine augmented data with the original to prevent degradation during continued training.

Our results are evaluated on the LibriSpeech-960h data, with and without shallow fusion [1], i.e., the integration of an external language model (LM) in the decoding step, where our method is able to reduce WER down to 2.55 and 6.27 on `test-clean` and `test-other`, respectively. We also conduct experiments on the ASR part of the CoVoST-2 data set for five languages, namely English, German, Catalan, French and Spanish, and show absolute improvements of up to 0.9 WER points.

2. RELATED WORK

Pseudo-labeling [2, 3, 4, 5] is an effective technique to use external models to generate new source-target training pairs from speech sources without transcriptions or target texts without audio. Examples are noisy student training [6], consistency training [7, 8, 9] and TTS-generated data [10, 11, 12]. Possible disadvantages of pseudo-labeling are its dependency on the quality of the data, and cost of integrating external models or tools, which is not necessary in our approach.

Other techniques generate new labeled data by assembling information solely from the existing training data. For example, MixSpeech [13] creates a new audio spectrogram by linearly interpolating two spectrograms. Our method creates new data instances by concatenation in the temporal dimension. This is similar to segmenting audio-target sequences into smaller paired units in the temporal dimension, for ASR [14, 15, 16] and speech-translation [17]. DA by segmentation requires an acoustic aligner, whereas our method does not rely on any external information.

Model	test-clean	w/shallow fusion	test-other	w/shallow fusion
Pre-trained	3.30	3.13	7.51	6.81
CT orig ∪ CatSelf	3.81	4.24	7.97	7.49
CT orig ∪ CatSpeaker	2.83 ±0.03	2.55 ±0.04	6.87 ±0.03	6.27 ±0.07
CT orig ∪ CatRandom	2.90 ±0.01	2.65 ±0.02	6.93 ±0.06	6.36 ±0.09

Table 1. Word Error Rate of *pre-trained* and *continued training (CT)* ASR models on LibriSpeech test-clean and test-other data sets with and without shallow fusion (SF). The "±" values indicate standard deviation over 3 runs.

Model	test (En)	test (De)	test (Ca)	test (Fr)	test (Es)
Pre-trained	19.76	20.47	13.64	15.41	14.66
CT orig ∪ CatSelf	20.75	26.06	14.18	16.05	15.21
CT orig ∪ CatSpeaker	19.67 ±0.00	19.71 ±0.02	12.79 ±0.18	14.98 ±0.00	14.05 ±0.04
CT orig ∪ CatRandom	19.63 ±0.13	19.55 ±0.04	12.89 ±0.02	15.04 ±0.07	14.13 ±0.05

Table 2. Word Error Rate of *pre-trained* and *continued training (CT)* ASR models trained on CoVoST-2 English (En), German (De), Catalan (Ca), French (Fr), and Spanish (Es) languages. The "±" values indicate standard deviation over 3 runs.

Similar concatenation-based techniques have been applied for special purposes, e.g., random audio concatenation in speech-to-speech translation [18], or generating longer inputs for document-level neural machine translation (NMT) [19]. Our work focuses on speech-to-text with the purpose of improving pre-trained models via continued training.

3. METHOD

Our DA strategy is to concatenate selected training instances in the temporal dimension, i.e., source-source and target-target concatenations. As there is no special separating token introduced by our method, we can make use of pre-trained off-the-shelf models. We evaluate three simple concatenation strategies: (1) CatSelf generates new training instances by repeating the original instance along the temporal dimension.(2) CatSpeaker makes use of speaker information and generates longer audio-text pairs spoken by the same person. (3) CatRandom generates new training instances by randomly concatenating audio-text pairs, spoken by different persons.

Our approach applied data augmentation *on-the-fly*. At the beginning of each epoch, we allow concatenations over the entire training data. Then, we combine the original training data and the augmented data, and apply length filtering before generating the training batches. By allowing concatenations over the entire training data instead of over only the current batch, we increase diversity of the augmented data. This concatenated data is then used for *continued training* of pre-trained models or training new models *from scratch*.

4. EXPERIMENTAL SETUP

4.1. Datasets and preprocessing

For the ASR tasks, we evaluate our method on LibriSpeech [20] and the CoVoST-2 [21] ASR dataset. For LibriSpeech, we combine the transcriptions in train960h and the extra 800M-word monolingual text data to train the LM. For CoVoST-2 ASR, we test on five languages: English (En), German (De), Catalan (Ca), French (Fr) and Spanish (Es). For the automatic speech translation (AST) tasks, we evaluate our method on CoVoST-2 and MuST-C for En-De. On both dataset, we use their own transcription-translation training data to train NMT models for knowledge distillation [22].

For all speech inputs, we extracted 80-dimensional log Mel-filterbank with 25ms FFT windows and 10ms frame shift. We filter instances with more than 3k frames. For transcriptions in LibriSpeech, we use the vocabulary file of 10k subword units from the FAIRSEQ GitHub repository[1]. For CoVoST-2 ASR tasks, we lowercased transcriptions and removed punctuation. For each language, we use 5k subword units. For translation tasks, we do not apply preprocessing on the translation data. For NMT and AST, the size of subword units are 5k and 8k for CoVoST-2 and MuST-C, respectively. All sub-word units are built using SentencePiece [23].

4.2. Model Architectures

We use FAIRSEQ [24, 25] for our implementation. For LibriSpeech, we used a pre-trained Transformer-based ASR model labeled *s_transformer_l* downloaded from the FAIRSEQ GitHub repository mentioned above. For shallow fusion, we use a Transformer-based LM of about 24M parameters. It has 6 layers with attention dimension of 512 and with FFN dimension of 2048. For CoVoST-2 ASR & AST and MuST-C AST, we used a Conformer architecture [26], labeled as *s2t_conformer*, of about 45M parameters. We follow the default configuration, with the exception of using 12 encoder layers and using attention type "attn-type=espnet". For NMT, we use a transformer of encoder-decoder-layers of size 3 and 6 for CoVoST-2 and MuST-C, respectively. Dimensions of attention and FFN-layer are 256 and 2048, respectively.

[1] https://github.com/facebookresearch/fairseq

Model	test (En)	test (De)	test (Ca)	test (Fr)	test (Es)
Pre-trained	19.64 ±0.09	20.40 ±0.07	13.58 ±0.09	15.35 ±0.05	14.79 ±0.18
FS orig ∪ CatSelf	21.59	21.47	13.67	16.47	15.53
FS orig ∪ CatSpeaker	19.65 ±0.04	19.50 ±0.05	12.09 ±0.19	14.87 ±0.11	14.03 ±0.11
FS orig ∪ CatRandom	19.44 ±0.17	19.22 ±0.00	11.96 ±0.12	14.94 ±0.08	14.14 ±0.06

Table 3. Word Error Rate of different ASR systems trained *from scratch (FS)* on the ASR part of CoVoST-2 English (En), German (De), Catalan (Ca), French (Fr) and Spanish (Es) languages. The "±" values are standard deviations over 3 runs.

4.3. Training and Inference

We use Adam optimizer [27] with inverse square root learning rate schedule for all experiments. For all experiments, we use a peak learning rate (lr) of 2e-3, with the exception of LM and NMT training where we use a lr of 5e-4 and of 1e-3, respectively. For pre-training and training from scratch, we adjust the warm-up steps for different settings. For continued training, we reset the optimizer with 1k warm-up. All speech-to-text experiments use a batch size of 40k×8 frames for training except for MuST-C, where we use 40k×2 and 25k×8 for ASR and AST, respectively. SpecAugment [28] is applied with a frequency mask of 27 and a time mask parameter of 100, with 2 masks along their respective dimension.

For LibriSpeech, we examined our strategies by training the pre-trained ASR model for 50k steps with validation step of 2k. The LM is trained for 200k steps with a batch size of 16k×2 tokens with 4k warm-up steps. For both ASR and LM, decoder-input and output embedding are shared.

For CoVoST-2 ASR, the pre-trained ASR models and the FC cases are trained for 30k steps, validated by every 500 steps. The exception is English which has more data. We thus train it for 60k steps with a validation step of 1k. All above models use 10k warm-up steps. For continued training, the En-ASR is trained for 20k steps, validated every 1k steps. De-ASR and Fr-ASR are trained for 10k steps whereas Ca-ASR and Es-ASR are trained for 8k steps. These four language pairs are validated every 500 steps. For AST, we initialize the encoder with a pre-trained En-ASR. The AST is then trained for 50k steps with 10k warm-up and validated every 1k steps.

For MuST-C, both ASR[2] and AST use 100k steps in training with 25k in warm-up and every 2k steps in validation. The NMT[3] is trained for 100 epochs with 8k warm-up steps, validated every epoch, and with a batch size of 100 sentences.

We use beam search of size 5 during inference. In shallow fusion, we use the last checkpoint with an interpolation weight of 0.3. For pre-training and training from scratch, we average the best 5 checkpoints by validation loss. For continued training, we average the *last* 5 checkpoints per validation step to prevent the averaging over pre-training checkpoints. For AST, we again average over the best 5 checkpoints.

[2]For both CoVoST-2 and MuST-C, the En-ASR models used in initialisation are trained on the original data only. In addition, the ASR models are obtained by averaging their 5 best checkpoints on their validation losses.

[3]CoVoST-2 NMT is similar except of a batch size of 16k tokens.

5. EXPERIMENTAL RESULTS

5.1. LibriSpeech

Table 1 lists Word-Error-Rate (WER) of the continued training experiments for each of the proposed concatenation strategies. CatSelf shows the worst performance in all settings and deteriorates even over the baseline model, resulting in WER degradation from 0.46 to 1.11. Both CatSpeaker and CatRandom, however, show significant improvements over the baseline system, with CatSpeaker performing slightly better than CatRandom throughout the experiments. We conjecture that speaker information is useful for ASR in the audiobooks domain, but the effect is very limited. Compared to the baseline that is trained on the original data only, CatSpeaker shows a reduction of 0.47 WER (14.2% relative) and of 0.64 WER (8.5% relative) on the `test-clean` and `test-other` splits, respectively. Further improvements can be achieved by using shallow fusion in decoding, resulting in 2.55 WER on `test-clean` (18.5% relative reduction) and 6.27 WER on `test-other` (7.9% relative reduction). All improvements over the pre-trained model are significant with $p < 0.005$ according to an approximate randomization test [29].

Table 4 shows an ablation study where training is continued using only augmented data without adding the original data. "CT orig" refers to continued training on the original training data set by the same number of updates as the augmented one. Here, we observed only minimal to no improvements. Continued training on the CatSelf data only shows largely worse performance compared to the baseline. A detailed inspection of the generated transcriptions reveals the underlying problem: The Transformer-based system tends to frequent repetitions in its output, a property that is introduced by this type of augmented data. Continued training on both CatSpeaker and CatRandom yields similar improvements with and without the inclusion of the original data.

5.2. CoVoST-2

Table 2 lists the WER of our concatenation strategies with continued training on 5 languages of the CoVoST-2 dataset. We see similar results to the LibriSpeech experiments, where CatSelf results in worse WER than the pre-trained models on all 5 languages. The degradation in WER ranges from 0.54 points for Catalan to 5.59 points for German, where the pre-trained systems is best for Catalan and worst for German.

Model	test-clean	with SF	test-other	with SF
Pre-trained	3.30	3.13	7.51	6.81
CT orig	3.26	3.05	7.38	6.82
CT CatSelf	41.29	46.48	54.31	57.80
CT CatSpeaker	2.94	2.55	7.09	6.42
CT CatRandom	2.94	2.64	7.31	6.51

Table 4. Ablation experiment: Word Error Rate of *continued training (CT)* using only original or augmented data on LibriSpeech `test-clean` and `test-other` data sets with and without shallow fusion (SF).

The CatSpeaker and CatRandom strategies yield similar WER improvements for each language. However, there is no consistent trend that might indicate if speaker information is useful or not. Throughout all languages, both CatSpeaker and CatRandom shows improvements over the pre-trained model with the largest WER improvement of 0.92 points (4.4% relative) for German, and the largest relative improvement in WER of 6.2% (0.75 points absolute) for Catalan. At the same time, the improvement on English is rather marginal even in the best case, i.e., 0.13 WER (0.7% relative) for CatRandom. We attribute this to the larger amount of the English training data compared to the other languages. The fact that this observation differs from the ASR improvements on LibriSpeech can be explained by the much simpler sentence complexity of the CoVoST-2 data. All improvements over the pre-trained model are significant with $p < 0.002$ except for English [29].

In Table 5 we repeat our ablation experiment to evaluate the contribution of the augmented data only. Similar to LibriSpeech, continued training using CatSelf data shows the worst performance compared to the baseline. An analysis of the transcriptions again reveals that the models tend to spurious repetitions in the output. In all cases except French, continued training using the original data also slightly degrades the model compared to the baseline. This is likely due to overfitting, as the pre-trained models use checkpoint-averaging to improve generalization, which is then reduced by continued training. Unlike the previous results on LibriSpeech, training on augmented data created by CatSpeaker and CatRandom mostly show worse performance over the pre-trained model. A slight improvement can be observed only for Catalan using the CatSpeaker data. We conjecture that the inclusion of the original data is vital for continued training on this dataset.

Model	test (En)	test (De)	test (Ca)	test (Fr)	test (Es)
Pre-trained	19.76	20.47	13.64	15.41	14.66
CT orig	20.10	20.90	13.98	15.38	15.31
CT CatSelf	117.87	118.20	110.32	114.20	112.15
CT CatSpeaker	27.36	22.34	12.94	15.69	15.22
CT CatRandom	25.46	21.74	13.68	15.69	15.54

Table 5. Ablation experiment: Word Error Rate *continued training (CT)* using only original or augmented data on CoVoST-2 English (En), German (De), Catalan (Ca), French (Fr), and Spanish (Es) languages.

Finally, we evaluate our concatenation strategies by training the entire ASR model from scratch for each language. Table 3 lists the results. For most cases, the improvements obtained by training from scratch are very close to those by continued training. Only for Catalan we observe further WER reduction of 0.86 compared to the continued training. Thus, our method also works for training from scratch if such training resources are available. Alternatively, one can use an off-the-shelf model and improve it via continued training with our method consuming much less computing power.

5.3. AST (En-De): MuST-C and CoVoST-2

We also evaluate our proposed DA strategies on two En-De speech-to-text translation tasks. Table 6 lists the chrF2 [30] scores of systems trained with "orig" (original data plus translations generated by knowledge distillation) and trained with combined data created by CatSpeaker or by CatRandom. Both concatenation strategies achieve significant improvements with $p < 0.00025$ both on MuST-C tst-COMMON and on CoVoST-2 test sets using the approximate randomization test implementation of SACREBLEU[4] [31].

Model	MuST-C tst-COMMON	CoVoST-2 test
orig	52.8 $\pm$0.0	47.65 $\pm$0.05
orig $\cup$ CatSpeaker	53.55 $\pm$0.05	48.55 $\pm$0.05
orig $\cup$ CatRandom	53.55 $\pm$0.05	48.45 $\pm$0.05

Table 6. chrF2 on MuST-C AST and CoVoST-2 AST (En-De). The "$\pm$" values indicate standard deviations over 2 runs.

The results show that our simple method is also applicable to AST where the speech-text alignments are not parallel.

6. CONCLUSION

We propose and evaluate temporal-concatenation as a data augmentation method for improving Transformer and Conformer based speech-to-text models. The method can be applied to improve pre-trained models without requiring extra information or external tools. We evaluate three concatenation strategies for ASR on LibriSpeech and CoVoST-2 data and found that concatenation by random and concatenation by speaker perform similarly and bring significant improvements. Finally, we evaluate our method for AST on Must-C and CoVoST-2 and also observed significant improvements. In the future, we would like to extend our method to other architectures and tasks.

7. ACKNOWLEDGEMENTS

This research was supported in part by the German research foundation DFG under grant RI-2221/4-1.

[4] nrefs:1|ar:10000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0

8. REFERENCES

[1] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N. Sainath, and Karen Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. 2018, pp. 369–375, IEEE.

[2] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Y. Hannun, Gabriel Synnaeve, and Ronan Collobert, "Iterative pseudo-labeling for speech recognition," in *INTERSPEECH*. 2020, pp. 1006–1010, ISCA.

[3] Yang Chen, Weiran Wang, and Chao Wang, "Semi-supervised ASR by end-to-end self-training," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 2787–2791, ISCA.

[4] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *CoRR*, vol. abs/2010.10504, 2020.

[5] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Self-training and pre-training are complementary for speech recognition," in *ICASSP*. 2021, pp. 3030–3034, IEEE.

[6] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le, "Improved noisy student training for automatic speech recognition," in *INTERSPEECH*. 2020, pp. 2817–2821, ISCA.

[7] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*. 2017, pp. 301–308, IEEE.

[8] Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramón Fernandez Astudillo, and Kazuya Takeda, "Back-translation-style data augmentation for end-to-end ASR," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*. 2018, pp. 426–433, IEEE.

[9] Takaaki Hori, Ramón Fernandez Astudillo, Tomoki Hayashi, Yu Zhang, Shinji Watanabe, and Jonathan Le Roux, "Cycle-consistency training for end-to-end speech recognition," in *ICASSP*. 2019, pp. 6271–6275, IEEE.

[10] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro J. Moreno, Yonghui Wu, and Zelin Wu, "Speech recognition with augmented synthesized speech," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*. 2019, pp. 996–1002, IEEE.

[11] Gary Wang, Andrew Rosenberg, Zhehuai Chen, Yu Zhang, Bhuvana Ramabhadran, Yonghui Wu, and Pedro J. Moreno, "Improving speech recognition using consistent predictions on synthesized speech," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 7029–7033, IEEE.

[12] Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Heiga Zen, Mohammadreza Ghodsi, Yinghui Huang, Jesse Emond, Gary Wang, Bhuvana Ramabhadran, and Pedro J. Moreno, "Semi-supervision in ASR: sequential mixmatch and factorized tts-based augmentation," in *INTERSPEECH*. 2021, pp. 736–740, ISCA.

[13] Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu, "Mixspeech: Data augmentation for low-resource automatic speech recognition," in *ICASSP*. 2021, pp. 7008–7012, IEEE.

[14] Thai-Son Nguyen, Sebastian Stüker, Jan Niehues, and Alex Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP*. 2020, pp. 7689–7693, IEEE.

[15] Lingxuan Ye, Gaofeng Cheng, Runyan Yang, Zehui Yang, Sanli Tian, Pengyuan Zhang, and Yonghong Yan, "Improving recognition of out-of-vocabulary words in E2E code-switching ASR by fusing speech generation methods," in *INTERSPEECH*. 2022, pp. 3163–3167, ISCA.

[16] Tsz Kin Lam, Mayumi Ohta, Shigehiko Schamoni, and Stefan Riezler, "On-the-fly aligned data augmentation for sequence-to-sequence ASR," in *INTERSPEECH*. 2021, pp. 1299–1303, ISCA.

[17] Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler, "Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation," in *Proceedings of ACL (2)*. 2022, pp. 245–254, Association for Computational Linguistics.

[18] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *ICML*. 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 10120–10134, PMLR.

[19] Toan Q. Nguyen, Kenton Murray, and David Chiang, "Data augmentation by concatenation for low-resource translation: A mystery and a solution," in *Proceedings of IWSLT*. 2021, pp. 287–293, Association for Computational Linguistics.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. 2015, pp. 5206–5210, IEEE.

[21] Changhan Wang, Juan Miguel Pino, Anne Wu, and Jiatao Gu, "Covost: A diverse multilingual speech-to-text translation corpus," in *Proceedings of LREC*. 2020, pp. 4197–4203, European Language Resources Association.

[22] Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe, "Source and target bidirectional knowledge distillation for end-to-end speech translation," in *Proceedings of NAACL-HLT*. 2021, pp. 1872–1881, Association for Computational Linguistics.

[23] Taku Kudo and John Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of EMNLP: System Demonstrations*. 2018, pp. 66–71, Association for Computational Linguistics.

[24] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT: Demonstrations*. 2019, pp. 48–53, Association for Computational Linguistics.

[25] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Miguel Pino, "Fairseq S2T: fast speech-to-text modeling with fairseq," in *Proceedings of AACL: System Demonstrations*. 2020, pp. 33–39, Association for Computational Linguistics.

[26] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*. 2020, pp. 5036–5040, ISCA.

[27] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR (Poster)*, 2015.

[28] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*. 2019, pp. 2613–2617, ISCA.

[29] Stefan Riezler and John T. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 2005, pp. 57–64, Association for Computational Linguistics.

[30] Maja Popovic, "chrf: character n-gram f-score for automatic MT evaluation," in *Proceedings of WMT*. 2015, pp. 392–395, Association for Computer Linguistics.

[31] Matt Post, "A call for clarity in reporting BLEU scores," in *Proceedings of WMT: Research Papers*. 2018, pp. 186–191, Association for Computational Linguistics.