Effective combination of pretrained models - KIT@IWSLT2022

Ngoc-Quan Pham[1] and **Tuan-Nam Nguyen**[1] and **Thai-Binh Nguyen**[1] and **Danni Liu**[2]
and **Carlos Mullov**[1] and **Jan Niehues**[1] and **Alexander Waibel**[1,3]
[1]Karlsruhe Institute of Technology
[2]Maastricht University
[3]Carnegie Mellon University, Pittsburgh PA, USA
firstname.lastname@kit.edu
firstname.lastname@maastrichtuniversity.nl

Abstract

Pretrained models in acoustic and textual modalities can potentially improve speech translation for both Cascade and End-to-end approaches. In this evaluation, we aim at empirically looking for the answer by using the wav2vec, mBART50 and DeltaLM models to improve text and speech translation models. The experiments showed that the presence of these models together with an advanced audio segmentation method results in an improvement over the previous End-to-end system by up to 7 BLEU points. More importantly, the experiments showed that given enough data and modeling capacity to overcome the training difficulty, we can outperform even very competitive Cascade systems. In our experiments, this gap can be as large as 2.0 BLEU points, the same gap that the Cascade often led over the years.

1 Introduction

Speech translation (ST) has been the main theme of IWSLT for more than a decade and it goes without saying between the traditional Cascade approach and the recent End-to-end (E2E) possibility, the former has always been preferred. Being able to divide the complicated ST to smaller sub-problems: automatic recognition, (often) re-segmentation (Cho et al., 2017) and machine translation, the cascade approach has the advantage of using more data to separately optimize the components. The E2E, on the other hand, relies on a single network architecture that requires an explicit speech-translation dataset.

Over the years of participation, we observed that the performance gap between E2E and cascade is reduced (Anastasopoulos et al., 2021), and there are three negative factors that outweigh the advantages of having a single architecture without the problem of error propagation (Sperber and Paulik, 2020).

- Data utilization: the end2end model can only be directly trained on parallel speech translation data, which is often lacking compared to speech-transcription or text translation data. Previously the SLT models would require a necessary pre-training step with ASR in order to have comparable results with cascade (Bansal et al., 2018; Pham et al., 2020c).

- Modeling power. The transition from shallow LSTM-based models (Sperber et al., 2019) to Transformer-based models (Pham et al., 2020a) resulted in a big leap in model capacity and showed the potential of the E2E approach.

- Better audio segmentation. Decoding directly from long audio files is infeasible due to the expensive memory requirement and the presence of other distractions such as breaks, noise or music. Applying either cascade or E2E models absolutely requires an audio segmentation step performed by a voice activity detection system. While the cascade systems can handle imprecise cuts based on a re-segmentation process (Cho et al., 2017), the E2E lacks this ability to recover from this training-testing condition mismatch.

In our work, we massively improved our end-to-end SLT systems for English→German with up to 6 BLEU points by directly addressing the aforementioned weaknesses:

- Pretrained acoustic (Baevski et al., 2020) and language models (Tang et al., 2020) are incorporated in modeling. This allowed for transferring the knowledge during the pretraining processes which contain a massive amount of data. This effect is further enhanced when combined with the pseudo labels generated by machine translation.

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 190 - 197
May 26-27, 2022 ©2022 Association for Computational Linguistics

- By using the pretrained models, we fully utilized the large architectures that improved the results further. More importantly, the pretrained acoustic model directly extracts features from audio waveforms which is potentially an advantage compared to the manually extracted features in the previous systems.
- The audio segmentation component is changed into a full neural-based solution combined with pretraining (Tsiamas et al., 2022). The new solution is not only more accurate, but also directly optimized on TED Talks giving the translation model more precise and complete segmentations compared to the generic voice activity detectors.

Moreover, we also applied the same techniques to improve the Speech Recognition and Machine Translation components of the Cascade system. They also benefit from the above factors, albeit to a limited extent. Unlike previous years when the Cascade was always the better performing system, for the first time we selected the E2E as our primary submission.

For the current evaluation campaign (Anastasopoulos et al., 2022), we also expanded the SLT systems for two new directions: English→Chinese and English→Japanese, with both of the approaches available. The resulting system is also used in a simultaneous setting located in the same evaluation campaign (Polák et al., 2022).

2 Data

Speech Corpora. For training and evaluation of our ASR models, we used Mozilla Common Voice v7.0 (Ardila et al., 2019), Europarl (Iranzo-Sánchez et al., 2020), How2 (Sanabria et al., 2018), Librispeech (Panayotov et al., 2015), MuST-C v1 (Di Gangi et al., 2019), MuST-C v2 (Cattoni et al., 2021) and Tedlium v3 (Hernandez et al., 2018) dataset. The data split is presented in the following table 1.

3 Cascade System for Offline Speech Translation

We address the offline speech translation task by two main approaches, namely cascade and end-to-end. In the cascaded condition, the ASR module (Section 3.1) receives audio inputs and generates raw transcripts, which will then pass through a Segmentation module (Section 3.2) to formulate well normalized inputs to our Machine Translation module (Section 3.3). The MT outputs are the final outputs of the cascade system. On the other hand, the end-to-end architecture is trained to directly translate English audio inputs into German text outputs (Section 3.4).

Table 1: Summary of the English data-sets used for speech recognition

Corpus	Utterances	Speech data [h]
A: Training Data		
Common Voice	1225k	1667
Europarl	33k	85
How2	217k	356
Librispeech	281k	963
MuST-C v1	230k	407
MuST-C v2	251k	482
TEDLIUM	268k	482
B: Test Data		
Tedlium	1155	2.6
Librispeech	2620	5.4

3.1 Speech Recognition

The speech recognition model is based on the wav2vec 2.0 architecture (Baevski et al., 2020) with a CTC decoder on top of the Transformer layers. The model is trained to output characters with a vocabulary of 30. Here we used the large version of Wav2vec 2.0 (24 hidden layers, hidden size is 1024), which was pre-trained on 53k hours of English audio data. The fine-tuning process used approximately 4.5k hours of audio (as illustrated in Table 1). The CTC decoder is supported by a 5-gram language model with a beam size of 100. The text corpus used to create the 5-gram model comes from the transcription label of the audio data.

3.2 Text Segmentation

The text segmentation in the cascaded pipeline serves as a normalization on the ASR output, which usually lacks punctuation marks and casing information. On the other hand, the machine translation system is often trained on well-written, high-quality bilingual data. Following the idea from (Nguyen et al., 2020), since punctuation and casing information always belong to words, we combine that info into 15 tags label (e.g **U. U, T! T$** ...). In which, punctuation has 5 types are **“. , ! ? $”** (**$** stands for no punctuation), casing information has 3 types are **“T”** (uppercase the first character of word), **“U”** (uppercase all character of word), **“L”** (lowercase all character of word). Our text segmentation model will become a sequence tagging model. We fine tune a BERT base-uncased model (Devlin et al., 2018) to predict tag label for each word in the input. Model has 12 hidden layers and hidden size is 768. The Yelp Review Dataset (Zhang et al., 2015) is used for training this model.

3.3 Machine Translation

For the machine translation module, we first re-use the English→German machine translation model from our last year' submission to IWSLT (Pham et al., 2020b). More than 40 millions sentence pairs being extracted from TED, EPPS, NC, CommonCrawl, ParaCrawl, Rapid and OpenSubtitles corpora were used for training the model. In addition, 26 millions sentence pairs are generated from the back-translation technique by a German→English translation system. A large transformer architecture was trained with Relative Attention. We adapted to the in-domain by fine-tuning on TED talk data with stricter regularizations. The same adapted model was trained on noised data synthesized from the same TED data. The final model is the ensemble of the two.

To fully use the available resources this year, we also fine-tune pretrained DeltaLM (Ma et al., 2021). We use the "base" configuration with 12 encoder and 6 decoder layers. Similar to the approach above, we conduct a two-step fine-tuning, first on WMT data and then on TED transcript-translation parallel data (except for English→Chinese where we directly fine-tuned on TED due to computation constraints). We also use this MT system to generate synthetic data from TEDLIUM transcripts for training the end-to-end systems.

For English→Japanese, the MT model based on DeltaLM and trained using 11.3M sentences from JESC, JParaCrawl, KFTT, TED and WikiMatrix datasets. Similar to the English→Chinese model, this model is also further finetuned on TED.

4 End-to-End System

4.1 Corpora

For training, we use all of the data available in Table 2. Here, the Speech Translation is pre-filtered using an ASR model to remove the samples that have a high mismatch between the manual label and transcription output[1].

Because of the multilingual condition, we combine the datasets for Japanese and Chinese from MuST-C, CoVoST (Wang et al., 2020) to train multilingual systems. Moreover, we followed the success of generating synthetic labels for audio utterances (Pham et al., 2020b) and translated the transcripts of TEDLIUM into all three languages using the MT models. This process required us to reconstruct the punctuations for the transcripts (Sperber and Paulik, 2020) and the translation in general is relatively noisy and incomplete (due the to fact that the segmentations are not necessarily aligned into grammatically correct sentences).

Table 2: Training data for E2E translation models.

Data	Utterances	Total time
English→German		
MuST-C v1	228K	408h
MuST-C v2	250K	408h
Europarl	32K	60h
Speech Translation	142K	160h
TEDLIUM	268K	415h
CoVoST	272K	424h
English→Japanese		
MuST-C v2	328K	420h
CoVoST	232K	400h
TEDLIUM	268K	415h
English→Chinese		
MuST-C	350K	480h
CoVoST	232K	400h
TEDLIUM	268K	415h

During training, the validation data is the Development set of the MuST-C corpus. The reason is that the SLT testsets often do not have the aligned audio and translation, while training end-to-end models often rely on perplexity for early stopping.

4.2 Modeling

In order to fully utilize the pretrained acoustic and language models, we constructed the SLT architecture with the encoder based on the wav2vec 2.0 (Baevski et al., 2020) and the decoder based on the autoregressive language model pretrained with mBART50 (Tang et al., 2020).

wav2vec 2.0 is a Transformer encoder model which receives raw waveforms as input and generates high level representations. The architecture consists of two main components: first a convolution-based *feature extractor* downsamples long audio waveforms into features that have similar lengths with spectrograms. After that, a deep

[1] Here we used BLEU score as the metric.

Transformer encoder uses self-attention and feed-forward neural network blocks to transform the features without further downsampling.

During the self-supervised training process, the network is trained with a constrastive learning strategy (Baevski et al., 2020), in which the features (after being downsampled) are randomly masked and the model learns to predict the quantized latent representation of the masked time step as well as encouraging the model to diversify the quantization codebooks by maximizing their entropies.

During the supervised learning step, we freeze the feature extraction weights to save memory since the first layers are among the largest ones and fine-tune all of the weights in the Transformer encoders. Moreover, in order to make the model more robust against the fluctuation in absolute positions when it comes to audio signals, as well as the training-testing mismatched condition happening when we have to use a segmentation model to find audio segments during testing, we added the relative position encodings (Dai et al., 2019; Pham et al., 2020a) to alleviate this problem.

Here we used the same pretrained model with the speech recognizer, with the large architecture pretrained with $53k$ hours of unlabeled data.

mBART50 is an encoder-decoder Transformer-based language model. During training, instead of the typical language modeling setting of predicting the next word in the sequence, this model is trained to reconstruct a sequence from its noisy version (Lewis et al., 2019) and later extended to a multilingual version (Liu et al., 2020; Tang et al., 2020) in which the corpora from multiple languages are combined during training. mBART50 is the version that is pretrained on 50 languages.

Architecture wise, this model follows the Transformer encoder and decoder (Vaswani et al., 2017). During fine-tuning, we can combine the mBART50 decoder with encoder pretrained with the wav2vec 2.0 so that each component contains the knowledge of one modality. The cross-attention layers connecting the decoder with the encoder are the parts that require extensive fine-tuning in this case, due to the modality mismatch between pretraining and finetuning.

Eventually, the model is easily extensible to a multilingual scenario by training on the combination of the datasets. The mBART50 vocabulary contains language tokens for all three languages and can be used to control the language output (Ha et al., 2016).

4.3 Speech segmentation

As pointed out in (Tsiamas et al., 2022), the quality of audio segmentation has a big impact on the performance of the speech translation models, which are trained on utterances corresponding to full sentences, often manually aligned, and this rarely happens with an automatic segmentation system.

With the advantage of neural architectures and pretrained models, we follow the SHAS method (Tsiamas et al., 2022) to train a Transformer-based audio segmentation model on the MuST-C v2 corpus. Based on the high-level audio features generated by wav2vec 2.0, the model predicts the probability of each frame belonging to an utterance or not with cross-entropy. Afterwards, given the probabilities of the frames in an audio sequence (which are actually averaged over several rolls for more consistent accuracy), a segmentation algorithm called probabilistic DAC is used to aggressively cut the segments at the points with lowest probabilities, and then trim the segments to get probabilities higher than a set threshold.

We found this method to be much more effective than other voice activity detectors such as WebRTC-VAD (Wiseman, 2016). In the next experimental part, it will be shown that the audio segmentation quality is one of the most important factors helping the E2E system. Here we closely followed the original implementations and parameters to obtain the neural segmenter.

5 Experimental Results

5.1 Speech Recognition

The quality of our ASR system is measured on two testsets: TEDLIUM and Librispeech (clean). For comparison, we also provide the WER from the models trained without pre-training, including the Transformers (Pham et al., 2019), Conformers (Gulati et al., 2020) and LSTMs (Nguyen et al., 2019).

Table 3: WER on Libri and TEDLIUM test sets.

Data	Libri	TEDLIUM
Conformer-based	3.0	4.8
Transformer-based	3.2	4.9
LSTM-based	2.6	**3.9**
wav2vec 2.0	**1.1**	4.2

It is notable that the latest ASR system with pretraining is substantially better than the same architecture (but with less layers) on both Librispeech and TEDLIUM tests. While the improvement on TEDLIUM is 12.5%, we observed a significant 63% improvement on Librispeech, which is enabled by the large amount of read speech included in pretraining. The wav2vec 2.0 layer is also considerably larger than both Transformer variants.

Compared with the LSTMs, the wav2vec model is 57% better in Librispeech, yet the former reaches lower error rate in TEDLIUM. Since TED Talks accounts for the majority of the training data, pretraining on a large amount of read speech might not fully transfer to a more formal and spontaneous speech style.

5.2 Machine Translation

In Table 4, we report the performance of the machine translation systems described in Section 3.3. We first show results for English-German when: 1) translating directly from the ground-truth transcripts, and 2) translating from the ASR outputs (Section 5.1).

First, we see incorporating the pretrained DeltaLM (Ma et al., 2021) improves translation quality from the ground-truth by 0.9-1.5 BLEU. The gain carries over to the speech translation performance when cascading with the ASR model, yet at a smaller scale of 0.5-0.8 BLEU. This suggests that the MT quality still degrades when coping with noisy inputs from ASR transcripts.

For Chinese and Japanese, the two newly added language in this year's evaluation campaign, we evaluate on the MuST-C tst-COMMON transcript-translation data. The BLEU scores are 28.3 and 19.5 respectively[2].

Table 4: Performance of the machine translation module in BLEU↑.

Testset en→de	**tst2015**	**tst2019**	**tst2020**
From ground-truth			
MT2021	33.9	28.5	32.3
MT2022	34.8	30.0	33.2
From ASR			
MT2021	26.1	25.1	27.9
MT2022	26.9	25.9	28.4

5.3 End-to-end Offline Speech Translation

Given two new factors coming into play for the End-to-end models, namely pretrained models and audio segmentation, the models are tested on the static test which is the tst-COMMON set from the MuST-C corpus (Di Gangi et al., 2019) with the pre-segmented utterances and labels. This testset is available for all three languages. The whole system is tested on the IWSLT testsets without utterance boundaries and labels are only provided in paragraphs (each talk is contained in one paragraph). In this condition, only English→German tests are available.

The results on this test for all three languages are presented in Table 5. On English-German, overall we managed to improve the purely supervised model with Transformers (Pham et al., 2020a) by 2.6 BLEU points. Using the pretrained weights from wav2vec and mBART is very effective for an additional 1.6 BLEU points, while we found that the relative attention also contributed for a 0.7 BLEU points, and training the model in the multilingual setting is also slightly better.

Table 5: BLEU scores on tst-COMMON from MuST-C

Model	BLEU
English-German	
E2E 2021	30.6
wav2vec + mBART	32.2
wav2vec + rel + mBART	32.9
wav2vec + rel + mBART + multi	33.2
English-Chinese	
wav2vec + rel + mBART + multi	24.5
English-Japanese	
wav2vec + rel + mBART + multi	16.9

Table 6: ST: Translation performance in BLEU↑ on IWSLT testsets (re-segmentation required). Progressive results from this year and last year end-to-end (E2E) and cascade (CD) are provided.

Testset →	**tst2015**	**tst2019**	**tst2020**
E2E2021	22.13	20.43	23.20
CD2021	24.95	21.07	25.4
E2E2021 + SHAS	26.66	24.55	25.58
+W2V-MBART	26.64	26.31	28.66
+REL	27.27	26.58	29.11
+MULTI	27.65	26.84	29.2
+ENSEMBLE	**27.87**	**27.61**	**30.05**
CD2022	26.84	25.91	28.35

[2] Using `tok.zh` and `tok.ja-mecab-0.996-IPA` respectively from sacreBLEU(Post, 2018)

The final results on previous IWSLT testsets are presented in Table 6. First of all, the new segmentation method SHAS managed to improve the translation results of our previous year's submission by up to 4.4 BLEU points (as can be see on tst2015 and tst2019). By using a stronger model with wav2vec and mBART pretrained modules, the results are vastly improved by 2.2 and 3.1 BLEU points on tst2019 and tst2020. The performance is incrementally improved even further, by combining different techniques including relative attention, multilingual training and ensemble. Eventually, we obtain a result which is 7.8 BLEU points better than the last year's end-to-end submission.

The cascade system is also improved this year, by using the pretrained ASR, MT and better segmentation. On tst2020, we managed to improve the BLEU score by 3 points. However this enhancement pales against the E2E, and this is our first participation in which the E2E convincingly outperformed the Cascade system.

6 Conclusion

If the end-to-end models remained as a promising approach in the previous evaluation campaigns, it eventually blooms as the superior solution when the conditions are met to overcome its problems, namely training difficulty, segmentation issues and inefficient data usage. While the performance of the E2E system is now better, we can still believe that its far from being practical given the size of the model and the required presence of an audio segmenter. Moreover, the Cascade system is still necessary since it can provide a distillation tool for the E2E, via pseudolabels for better data utilization. The development of both approaches remains to be interesting awaiting the future achievement in multilingual and multimodal unsupervised and self-supervised training.

Acknowledgments

The projects on which this paper is based were funded by the Federal Ministry of Education and Research (BMBF) of Germany under the numbers 01IS18040A. The authors are responsible for the content of this publication.

References

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z. Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, et al. 2021. Findings of the iwslt 2021 evaluation campaign. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29.

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.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

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.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017. NMT-based segmentation and punctuation insertion for real-time spoken language translation. In *Interspeech 2017*. ISCA.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Thai Binh Nguyen, Quang Minh Nguyen, Thi Thu Hien Nguyen, Quoc Truong Do, and Chi Mai Luong. 2020. Improving Vietnamese Named Entity Recognition from Speech Using Word Capitalization and Punctuation Recovery Models. In *Proc. Interspeech 2020*, pages 4263–4267.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2019. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. *arXiv preprint arXiv:1910.13296*.

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.

Ngoc-Quan Pham, Thanh-Le Ha, Tuan-Nam Nguyen, Thai-Son Nguyen, Elizabeth Salesky, Sebastian Stüker, Jan Niehues, and Alex Waibel. 2020a. Relative Positional Encoding for Speech Recognition and Direct Translation. In *Proc. Interspeech 2020*, pages 31–35.

Ngoc-Quan Pham, Thai-Son Nguyen, Thanh-Le Ha, Tuan-Nam Nguyen, Maximilian Awiszus, Felix Schneider, Sebastian Stüker, and Alexander Waibel. 2020b. Tkit's iwslt 2020 slt translation system. In *Proceedings of the 17th International Workshop on Spoken Language Translation (IWSLT 2020)*.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Muller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377*.

Ngoc-Quan Pham, Felix Schneider, Tuan Nam Nguyen, Thanh-Le Ha, Thai-Son Nguyen, Maximilian Awiszus, Sebastian Stüker, and Alex Waibel. 2020c. Kit's iwslt 2020 slt translation system. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 55–61.

Peter Polák, Ngoc-Quan Ngoc, Tuan-Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. 2022. CUNI-KIT System for Smultaneous Speech Translation Task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *arXiv preprint arXiv:1904.07209*.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. *arXiv preprint arXiv:2004.06358*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

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

John Wiseman. 2016. python-webrtcvad. https://github.com/wiseman/py-webrtcvad.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification . *arXiv:1509.01626 [cs]*.