# KIT's IWSLT 2021 Offline Speech Translation System

**Tuan-Nam Nguyen, Thai-Son Nguyen, Christian Huber, Maximilian Awiszus,**
**Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, Sebastian Stüker, Alexander Waibel**
Karlsruhe Institute of Technology
firstname.lastname@kit.edu

## Abstract

This paper describes KIT'submission to the IWSLT 2021 Offline Speech Translation Task. We describe a system in both cascaded condition and end-to-end condition. In the cascaded condition, we investigated different end-to-end architectures for the speech recognition module. For the text segmentation module, we trained a small transformer-based model on high-quality monolingual data. For the translation module, our last year's neural machine translation model was reused. In the end-to-end condition, we improved our Speech Relative Transformer architecture to reach or even surpass the result of the cascade system.

## 1 Introduction

As in previous years, the cascade system's pipeline is constituted by an ASR module, a text segmentation module and a machine translation module. In this year's evaluation campaign, we investigated only sequence-to-sequence ASR models with three architectures. The segmentation module is basically a monolingual system which translates a disfluent, broken, uncased text (i.e. ASR outputs) into a more fluent, written-style text with punctuations in order to match the data conditions of the translation system. The machine translation module's architecture is the same as the previous year's. For the end-to-end system, we improved from our last year's Speech Relative Transformer architecture (Pham et al., 2020a). As a result, the end-to-end system can produce better results on certain test sets and approach the performance on some others compared to the cascade system this year, while the end-to-end system was the dominant approach last year.

The rest of the paper is organized as followed. Section 2 describes the data set used to train and test the system. It is then followed by Section 3 providing the description and experimental results of both the cascade and the end-to-end system. In the end, we conclude the paper with Section 4.

## 2 Data

**Speech Corpora.** For training and evaluation of our ASR models, we used Mozilla Common Voice v6.1 (Ardila et al., 2019), Europarl (Koehn, 2005), 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.

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

| Corpus | Utterances | Speech data [h] |
|---|---|---|
| **A: Training Data** | | |
| Mozilla 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 |

**Text Corpora.** We collected the text parallel training data as presented in Table 2.

## 3 Offline Speech Translation

We address the offline speech translation task by two main approaches, namely cascade and end-to-end. In the cascade 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,

*Proceedings of the 18th International Conference on Spoken Language Translation*, pages 125–130
Bangkok, Thailand (Online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Table 2: Text Training Data

| Dataset | Sentences |
|---|---|
| TED Talks (TED) | 220K |
| Europarl (EPPS) | 2.2MK |
| CommonCrawl | 2.1M |
| Rapid | 1.21M |
| ParaCrawl | 25.1M |
| OpenSubtitles | 12.6M |
| WikiTitle | 423K |
| Back-translated News | 26M |

the end-to-end architecture is trained to directly translate English audio inputs into German text outputs (Section 3.4).

### 3.1 Speech Recognition

**Data preparation and Segmentation tool** After collecting all audios from all data sets mentioned in Section 2, we calculated 40 features of Mel-filterbank coefficients for ASR training. To generate labels for the sequence-to-sequence ASR models, we used the Sentence-Piece toolkit (Kudo and Richardson, 2018) to train 4000 different byte-pair-encoding (BPE). The WerRTCVAD toolkit (Wiseman, 2016) was used to segment the audio in the testing phase.

**Model** As in previous years (Pham et al., 2019a, 2020b), we used only sequence-to-sequence ASR models, which are based on three different network architectures: The long short-term memory (LSTM), the Transformer and the Conformer. LSTM-based models (Nguyen et al., 2020) consist of 6 bidirectional layers for the encoder and 2 unidirectional layers for the decoder, both encoder and decoder layers have 1536 units. The Transformer-based models presented in (Pham et al., 2019b) have 24 layers for the encoder and 8 layers for the decoder. The Conformer-based models (Gulati et al., 2020) comprise 16 layers for the encoder and 6 layers for the decoder. In both the Transformer-based and the Conformer-based models, the size of each layer is 512 and the size of the hidden state in the feed-forward sublayer is 2048. The speech data augmentation technique was used to reduce overfitting as described in (Nguyen et al., 2020). In order to train a deep network effectively, we also applied Stochastic Layers (Pham et al., 2019b) with a dropping layer rate of 0.5 on both Transformer-based and Conformer-based models.

### 3.2 Text Segmentation

The text segmentation in the cascaded pipeline serves as a normalization on the ASR output, which usually lacks punctuation marks, proper sentence boundaries and reliable casing. On the other hand, the machine translation system is often trained on well-written, high-quality bilingual data. Following the idea from (Sperber et al., 2018a), we build the segmentation as a monolingual translation system, which translates from lower-cased, without-punctuation texts into texts with case information and punctuation, prior to the machine translation module.

The monolingual translation for text segmentation is implemented using our neural speech translation framework `NMTGMinor`[1](Pham et al., 2020a). It is a small transformer architecture, consisting of a 4-layer encoder and 4-layer decoder, in which each layer' size is 512, while the inner size of feed-forward network inside each layer is 2048. The encoder and decode are self-attention blocks, which have 4 parallel attention heads. The training data for that are the English part extracted from available multilingual corpora: EPPS, NC, Global Voices and TED talks. We trained the model for 10 epochs, then we fine-tuned it on the TED corpus for 30 epochs more with stronger drop-out rate. Furthermore, to simulate possible errors in the ASR outputs, a similar model is trained on artificial noisy data and the final model is the ensemble of the two models.

The trained model is then utilized to translate the ASR outputs in a shifting window manner and the decisions are drawn by a simple voting mechanism. For more details, please refer to (Sperber et al., 2018a).

### 3.3 Machine Translation

For the machine translation module, we 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

[1] https://github.com/quanpn90/NMTGMinor

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.

### 3.4 End-to-End Model

**Corpora** This year, the training data consists of the second version of the MUST-C corpus (Di Gangi et al., 2019), the Europarl corpus (Iranzo-Sánchez et al., 2020), the Speech Translation corpus and the CoVoST-2 (Wang et al., 2020) corpus provided by the organizer. The speech features are generated with the in-house Janus Recognition Toolkit. The ST dataset is handled with an additional filtering step using an English speech recognizer (trained with the its transcripts with the additional Tedlium-3 training data).

Following the success of generating synthetic audio utterances, the transcripts in the Tedlium-3 corpus are translated into German using the cascade built in the previous year's submission (Pham et al., 2020b). In brief, the translation process required us to preserve the audio-text alignment from the original data collection and segmentation process. As a results, we used the Transformer-based punctuation inserting system from IWSLT2018 (Sperber et al., 2018b) to reconstruct the punctuations for the transcripts followed by the translation process that preserves the same segmentation information. Compared to the human translation from the speech translation datasets, this translation is relative noisier and incomplete (due to the segmentations are not necessarily aligned with grammatically correct sentences).

The end result of the filtering and synthetic creation process is the complete translation set, as summarised in Table 3

Table 3: Training data for E2E translation models.

| Data | Utterances | Total time |
|---|---:|---:|
| MuST-C | 229K | 408h |
| Europarl | 32K | 60h |
| Speech Translation | 142K | 160h |
| Tedlium-3 | 268K | 415h |
| CoVoST | 288K | 424h |

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.

**Modeling** The main architecture is the deep Transformer (Vaswani et al., 2017) with stochastic layers (Pham et al., 2019b). The encoder self attention layer uses Bidirectional relative attention (Pham et al., 2020a) which models the relative distance between one position and other positions in the sequence. This modeling is bidirectional because the distance is distinguished for each direction from the perspective of one particular position. The main models use a "Big" configuration with 16 encoder layers and 6 decoder layers, and they are randomly dropped in training according to the linear schedule presented in the original work, where the top layer has the highest dropout rate $p = 0.5$. The model size of each layer is 1024 and the inner size is 4096. We experimented with different activation functions including GELU (Hendrycks and Gimpel, 2016), SiLU (Elfwing et al., 2018) and the gated variants similar to the gated linear units (Dauphin et al., 2017). Also, each transformer block (encoder and decoder) is equipped with another feed-forward neural network in the beginning (Lu et al., 2019). Our preliminary experiments showed that GeLU and SiLU provided a slightly better performance than ReLU, and our final model is the ensemble of the three configurations that are identical except the activation functions.

First, the encoders are pretrained using the data portions containing English texts to make training SLT stable. With the initialized encoder, the networks can be trained with an aggressive learning rate with 4096 warm-up steps. Label-smoothing and dropout rates are set at 0.1 and 0.3 respectively for all models. Furthermore, all speech inputs are augmented with spectral augmentation (Park et al., 2019; Bahar et al., 2019). All models are trained for 200000 steps, each consists of accumulated 360000 audio frames. Using the model setup like above, we managed to fit a batch size of around 16000 frames to 24 GB of GPU memory.

**Speech segmentation** As reflected from last year's experiments, audio segmentation plays an important role in the performance of the whole system, and the end-to-end model unfortunately does not have control of segmentation, as it is a prerequisite before training one. During evaluation, we relied on the WerRTCVAD toolkit (Wiseman, 2016) to cut the long audio files into segments of reasonable length, and the tool is also able to rule out silence and events that do not belong to human speech, such as noise and music.

Overall, we improved the submission from last year (Pham et al., 2020b) using stronger models together with a more accurate segmentation tool.

## 3.5 Experimental Results

### 3.5.1 Cascade Offline Speech Translation

**Speech Recognition.** We tested our ASR systems on two datasets, Tedlium and Libri test set. The ensemble of LSTM-based and Conformer-based sequence-to-sequence model provide the best results, which are 2.4 and 3.9 WERs respectively for two test set Table 4.

Table 4: WER on Libri and Tedlium sets

| Data | Libri | Tedlium |
|---|---|---|
| Conformer-based | 3.0 | 4.8 |
| Transformer-based | 3.2 | 4.9 |
| LSTM-based | **2.6** | **3.9** |
| Ensemble | **2.4** | 3.9 |

**Machine Translation.** We do not train any new machine translation module but re-use last year's model, thus, we do not conduct experiments and comparisons with different machine translation systems. We submitted one cascased model with our audio segmentation.

### 3.5.2 End-to-end Offline Speech Translation

Our models are tested on two different setups. On the one hand, we evaluated the model on the tst-COMMON (2nd version) of the MuST-C corpora. Due to the incompatibility between the models and the audio data that requires resegmentation, we rely on the dev and test sets of MuST-C to evaluate the ability to translate on "ideal" conditions. As mentioned above, our ensemble managed to reach 32.4 BLEU points on this test set[2].

On the other hand, we used the testsets from 2010 to 2015 to measure the progress from last year in the condition requiring audio segmentation. In this particular comparison as shown in Table 5, we showed that using a stronger model together with better voice detection not only improves the SLT results by up to 1.9 BLEU points (in *tst2014*) but also outperforms the strong cascade in 2 different sets: *tst2013* and *tst2014*, in which the difference could be even 1 BLEU point. There is still a performance gap in the last two tests, however, a strong E2E system can now trade blow with a strongly tuned cascade. The deciding factor, in our opinion, is audio segmentation because this is the sole advantage of the cascade which can recover from badly cut segments[3].

Table 5: ST: Translation performance in BLEU↑ on IWSLT testsets (re-segmentation required). Progressive results from this year and last year end-to-end (E2E) and cascades (CD) are provided.

| *Testset* → | CD 2020 | E2E 2020 | E2E 2021 |
|---|---|---|---|
| tst2010 | **26.68** | 24.27 | 25.28 |
| tst2013 | 28.60 | 28.13 | **29.62** |
| tst2014 | 25.64 | 25.46 | **27.32** |
| tst2015 | **24.95** | 21.82 | 22.13 |

## 4 Conclusion

In this year's evaluation campaign, the end-to-end model proves to be a very promising approach since it can compete or even transcend the best cascade model in offline speech translation task. As a note for future work, we would like to investigate two-stage speech translation models (Sperber et al., 2019) using transformer architectures and compare them with our recent speech translation end-to-end models.

## Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n°825460 and the Federal Ministry of Education and Research (Germany)/DLR Projekt-tträger Bereich Gesundheit under grant agreement n° 01EF1803B.

## References

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*.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. *arXiv preprint arXiv:1911.08876*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021.

[2] Unfortunately the comparison to last year tst-COMMON (30.6 is not available due to version mismatch.

[3] Changing the VAD parameters does not affect the performance of the cascade significantly, while the E2E can be badly afffected

Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

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)*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11.

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 recognitio. In *Proc. Interspeech 2020*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Ted-lium 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-Cerdà, Javier Jorge, Nahuel Roselló, Adrià 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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

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.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. 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.

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. *arXiv preprint arXiv:1904.08779*.

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, Juan Hussain, Felix Schneider, Jan Niehues, Sebastian Stüker, and Alexander Waibel. 2019a. The iwslt 2019 kit speech translation system. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Muller, and Alex Waibel. 2019b. 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 Alexander Waibel. 2020b. Kit's iwslt 2020 slt translation system. In *Proceedings of the 17th International Workshop on Spoken Language Translation (IWSLT 2020)*.

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. In *Proc. ACL 2019*.

Matthias Sperber, Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Thanh-Le Ha, Sebastian Stüker, and Alex Waibel. 2018a. KIT's IWSLT 2018 SLT Translation System. In *15th International Workshop on Spoken Language Translation 2018*. IWSLT.

Matthias Sperber, Ngoc Quan Pham, Thai Son Nguyen, Jan Niehues, Markus Müller, Thanh-Le Ha, Sebastian Stüker, and Alex Waibel. 2018b. KIT's IWSLT 2018 SLT Translation System. In *"Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)"*, Brussels, Belgium.

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: A massively multilingual speech-to-text translation corpus.

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