1 End-to-End ST

1.1 Dataset

The dataset used is a subset of the CoVoST2 dataset. The CoVoST2 is a large-scale multilingual speech translation corpus with 21 languages into English and 15 languages from English [Wang(2020a)]. For our training, we used the part of the dataset with translations from English to German.

The split of CoVoST2 for train, dev and test set for the translation direction English to German is shown in the following Table [Wang(2020b)].

English to German								
Hours			Speakers			Src./Tgt. Tokens		
Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
364	26	25	10k	4k	9k	3M/3M	156K/155K	4M/4M

Table 1: CoVoST2 number and split of data for English to German ST [Wang(2020b)]

1.2 Model Architecture

As the model architecture, we used a Speech-to-Text Transformer with an encoder and a decoder.

Encoder:

- The encoder has 12 layers

Decoder

- The decoder has 6 layers

Common Features of Encoder and Decoder:

- They use dropout for regularization in multiple locations
- They use LayerNorm for normalization in various parts
- They use Multihead Attention and Linear layers
- They use SinusoidalPositionalEmbedding for positional embeddings

1.3 Performance

The training lasted for 2 days, during which 81 epochs were computed. In the following table, the performance of end-to-end ST on the dev and test sets is shown:

Type	Beam Size	Score Type	Score Value
Dev Set	5	Loss	3.57
Test Set	5	BLEU	5.73

1.3.1 Comparison

In the following table, cascaded ST and end-to-end ST are compared regarding the BLEU score:

Model	Beam size	Score Type	Score Result
ASR	5	WER	9.38
MT	4	BLEU	28.7
Cascaded	5 and 4	BLEU	18.6
End-to-End	5	BLEU	5.73

Explanation of Performance: At present, the cascaded ST model, with a BLEU score of 18.6, significantly outperforms the end-to-end ST model, which has a BLEU score of only 5.73.

A potential reason for the relatively poor performance of the end-to-end ST model is that we initially started the training with a very small dataset and then continued on a larger dataset, building upon the previous checkpoints from the very small dataset. This approach likely led to relatively slow and poor training performance.

The more critical point, however, is that we always started the training with new Byte Pair Encoding (BPE), as we unfortunately had to redo the entire preprocessing at the start of each training session. This possibly resulted in us having BPEs of the same size, but with completely different tokens.

To address these two issues, we have now initiated training directly with the large dataset and consistently used the same BPE. However, this training has been in the queue for quite some time. We will provide the results from this approach as soon as they are available, in case it leads to any improvements.

References

[Wang(2020a)] Changhan Wang. Covost 2 and massively multilingual speech-to-text translation, 2020a. URL `https://arxiv.org/abs/2007.10310`.

[Wang(2020b)] Changhan Wang. Covost: A large-scale multilingual speech-to-text translation corpus, 2020b. URL `https://github.com/facebookresearch/covost`.