End-to-End Speech Translation with Cross-lingual Transfer Learning

Shuichiro Shimizu[1] Chenhui Chu[1] Sheng Li[2] Sadao Kurohashi[1]
[1]Kyoto University, Kyoto, Japan
[2]National Institute of Information and Communications Technology, Kyoto, Japan
{sshimizu, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp
sheng.li@nict.go.jp

1 Introduction

Speech Translation (ST) is the task of translating input speech into translated text [1]. There are mainly two approaches for this task: cascaded approach, where Automatic Speech Recognition (ASR) and Machine Translation (MT) are chained together, and end-to-end approach, where a single sequence-to-sequence model directly translates between audio signals and target text.

Cascaded approach has a problem of error propagation: possible errors produced by ASR are input to MT without any fix. On the other hand, end-to-end approach does not produce such kind of ASR errors, because the target text is directly produced. Therefore, end-to-end approach is becoming more popular in recent days.

However, end-to-end approach still has a problem of data scarcity. It is not easy to collect ST datasets, because they require triplets of source speech, source transcript, and target translation. To address the problem of data scarcity, previous studies have tried to use information from other languages. One of the existing models is the joint multilingual model, where a single model is used to process multiple languages [2, 3, 4] (Figure 1, left). Unfortunately, this joint model only improves performance on high-resource language pairs (Table 2). We hypothesize that this is because of 1) the difference in data size for each language and 2) pushing too many tasks on a single model. These problems are thoroughly examined in the field of multilingual MT [5, 6, 7].

In the meantime, transfer learning for ST using ASR has recently been explored [8, 9]. They show that ASR pre-training on a high resource language can improve low-resource ST performance (Figure 1, right). Another direction of improving ASR using ST has also been explored [10]. Their studies are based on the assumption that both ASR and ST can be viewed as what is called a speech-to-text task, and one can improve the other by transfer learning.

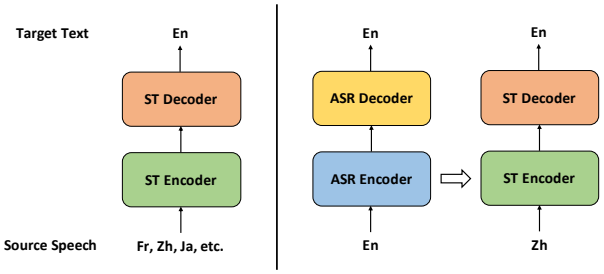

Figure 1 Existing methods for multilingual ST. Left: joint model. Right: ST with ASR pre-training.

However, ST is a more complicated task than ASR, because it requires more abstract representations to perform translation. There is a study that shows the complexity of ST encoder [11]. On the decoder side, some studies have tried to use MT decoder to initialize ST decoder [4, 12]. These studies are showing improvements, but the task difference between ASR encoder and ST encoder, and between MT decoder and ST decoder leaves room for further improvements.

To address the above problems, we propose cross-lingual transfer learning for end-to-end ST, where we transfer parameters of ST models from one language pair to another. At most three languages are involved during training, and transfer of information is performed on the same task. This can solve the problem of too many tasks on a single model, and also the task difference between ASR/MT and ST. Our method shows improvement up to 2.28 BLEU, and is effective for both high-resource and low-resource language settings.

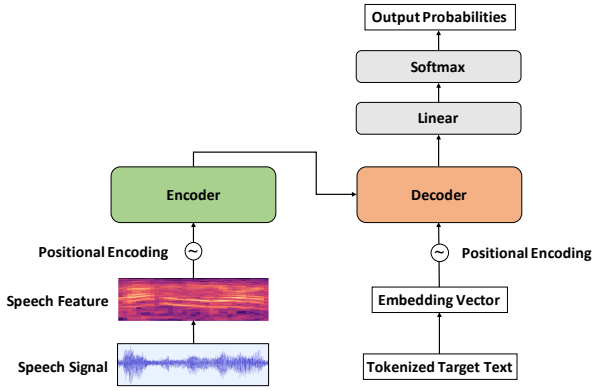

Figure 2 Overview of a speech-to-text task.

2 Preliminary

End-to-end ST is one kind of speech-to-text tasks (ASR/ST). Figure 2 shows an overview of the task.

In a speech-to-text task, we have a pair of source speech and target text, denoted as $\mathcal{S} = \{(\boldsymbol{x}, \boldsymbol{y})\}$. $\boldsymbol{x}$ and $\boldsymbol{y}$ are from the same language if the task is ASR, and different if the task is ST. Here, $\boldsymbol{x} = (x_1, \ldots, x_T)$ is a sequence of acoustic signals per utterance, and $\boldsymbol{y} = (y_1, \ldots y_N)$ is a sequence of characters per utterance. Each utterance mostly corresponds to one sentence, but can be two or more sentences.

First, we extract speech features (i.e., filterbank) from audio signals and get a sequence of frames. Fourier transform, Mel-scale conversion, log conversion are sequentially applied to the frames, which becomes a set of d dimensional vectors called d-dimensional log Mel filterbank.

Then we input the vector into the encoder of a sequence to sequence architecture. Here we use Transformer [13]. We apply positional encoding to the input vectors, which are fed into the encoder. Each encoder layer has a self-attention layer and a feed-forward layer with addition & normalization layers. On the decoder side, embedded vectors are calculated and positional encoding is applied as well. Each decoder layer has a masked self-attention layer, a cross-attention layer and a feed-forward layer with addition & normalization layers. After decoder layers, we have a linear layer and a softmax layer. The model is updated using cross-entropy loss:

$$L = -\log P(\boldsymbol{y}|\boldsymbol{x})$$

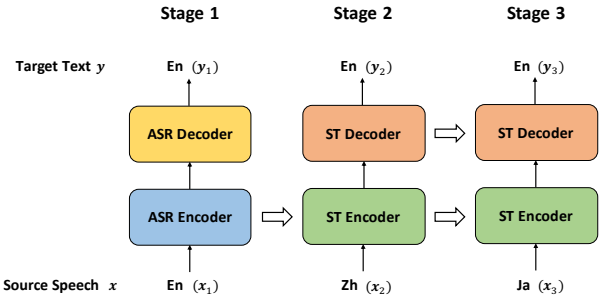

Figure 3 Proposed method. ST with ST pre-training and ASR encoder pre-training.

3 Cross-lingual Transfer for ST

The overview of our method is described in Figure 3. We use 3 speech-text pairs, $\mathcal{S}_1 = \{(\boldsymbol{x}_1, \boldsymbol{y}_1)\}$, $\mathcal{S}_2 = \{(\boldsymbol{x}_2, \boldsymbol{y}_2)\}$, and $\mathcal{S}_3 = \{(\boldsymbol{x}_3, \boldsymbol{y}_3)\}$. Source speech $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ are from different languages, and target text $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$ are all English. There is no overlap between $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$.

Stage 1: ASR pre-training First, we trained an ASR model using $\mathcal{S}_1 = \{(\boldsymbol{x}_1, \boldsymbol{y}_1)\}$ (En-En in Figure 3) to initialize ST encoder, following previous work [3, 8].

Stage 2: ST pre-training Secondly, we trained ST on one language pair using $\mathcal{S}_2 = \{(\boldsymbol{x}_2, \boldsymbol{y}_2)\}$ (Zh-En in Figure 3). We call the source language $\boldsymbol{x}_2$ as the "second language." As for encoder initialization, all self-attention layers, feed forward layers, and addition & normalization layers are shared with the ASR encoder at stage 1.

Stage 3: ST fine-tuning Finally, using the ST encoder and decoder as initialization, we trained ST on another language pair using $\mathcal{S}_3 = \{(\boldsymbol{x}_3, \boldsymbol{y}_3)\}$ (Ja-En in Figure 3). We call the source language $\boldsymbol{x}_3$ as the "third language." As for initialization, all of the encoder layers, the decoder layers, the linear layer, and the softmax layer are shared with those at stage 2.

4 Experiments

4.1 Dataset

We used CoVoST2 [2], a large-scale multilingual ST corpus which covers translations from 21 languages to English and from English to 15 languages. We used the former part of 21 languages to English, and chose 13 languages in addition to English (Table 1) following previous work [3]. We can see that English, French, Germany, and Spanish are high-resource languages, while others are low-resource ones.

Table 1 Number of utterances per language. Empty audio files included in the original dataset are removed during pre-processing.

	En	Fr	De	Es	Zh	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
Train	289,413	207,372	127,824	79,013	7,085	3,966	2,283	2,160	2,337	1,843	1,358	1,119	1,243	1,241
Dev.	15,531	14,760	13,511	13,221	4,843	1,624	1,758	1,349	1,125	509	384	635	792	690
Test	15,530	14,760	13,511	13,221	4,898	1,629	1,695	1,595	1,629	360	786	684	844	690

4.2 Pre-processing

We extracted 80-dimensional log Mel filterbank features with frames of 25 ms length and 10 ms stride. Utterance level cepstral mean variance normalization (CMVN) was applied to the feature. SpecAugment [14] with LB policy was applied to avoid overfitting. SentencePiece [15] was used to make character vocabulary for each language.

4.3 Model training

We used the fairseq S2T [3] toolkit for the experiments. We used Transformer with 6 encoder layers and 12 decoder layers where the hidden dimension size is 256 and the number of attention heads is 4. Label smoothing was applied when computing the cross-entropy loss.

Stage 1: ASR pre-training First, we conducted English ASR pre-training, because English is the highest resource language in this corpus. We used the adam optimizer with an inverse square root learning rate scheduler. We linearly increased the learning rate from 0 to 0.001 until 10,000 updates. Thereafter we decayed the learning rate proportional to the number of updates. We stopped training after 100,000 updates and averaged the parameters over the last 10 epochs.

Stage 2: ST pre-training Secondly, we loaded the parameters from the ASR encoder at stage 1, and trained ST on a language pair $\mathcal{S}_2 = \{(\boldsymbol{x}_2, \boldsymbol{y}_2)\}$. As source (second) languages $\boldsymbol{x}_2$, we chose French and Chinese. We chose French because it is the highest resource language pair, and Chinese because it is the highest one other than European languages. Again, we used the adam optimizer with an inverse square root learning rate scheduler, this time the learning rate was increased to 0.002 and then decayed. We stopped training after we had seen no improvement in terms of development set loss over 10 epochs, and used the best epoch for evaluation.

Stage 3: ST fine-tuning Finally, we loaded all the parameters at stage 2, and trained ST on another language pair $\mathcal{S}_3 = \{(\boldsymbol{x}_3, \boldsymbol{y}_3)\}$. The vocabulary was also shared with the second language. The optimizer and the learning rate scheduling was the same as stage 2. The model was trained for the same epochs as stage 2 and the last checkpoint was used for evaluation.

4.4 Decoding and Evaluation

For decoding, beam search was used with beam size 5. BLUE score was calculated with sacreBLEU [16].

4.5 Results

Table 2 shows the BLEU scores for test sets. The first row is our baseline, which is ST with ASR encoder pre-training. The second and third rows show the results of our proposed method. Cross-lingual Fr denotes our proposed method where the second language is French, and cross-lingual Zh is a Chinese one. These are based on our experiments, and the last two rows are from a previous study [3]. They show that the joint model improves the results compared to bilingual one in Fr, De, Es, and Zh, but in other languages the performance is degrading.

In our experiments, the scores of either of our methods outperformed the baseline scores in languages other than Chinese. For example, in Spanish-English ST, our cross-lingual Fr is better than bilingual one by 2.28 BLEU.

We conjecture that the linguistic similarity between the second language and the third language is important in our method. For example, when the second language is French, Ja-En ST BLEU score is 0.24, which is lower than the bilingual one. However, when the second language is Chinese, it improves to 0.94. French and Japanese have little in common in terms of vocabulary and grammar, but Chinese and Japanese have similar vocabularies.

4.6 Case Analysis

Table 3 shows an example from De-En, which is a high-resource language pair, and Ja-En ST, which is a low-resource one.

In De-En ST, cross-lingual Fr produced the best translation. Bilingual one could not correctly predict the position of *From the earth*. Cross-lingual Zh could predict the position correctly, but it is *Of the earth*, not *From the earth*. We

Table 2 BLEU scores for test set. *Encoder initialized with English ASR. †Wang et al., 2020 [3]. ‡Trained jointly on 21 languages with temperature based sampling ($T = 2$).

	Fr	De	Es	Zh	Tr	Ar	Sv	Lv	Sl	Ta	Ja	Id	Cy
Bilingual*	24.66	14.12	20.65	**4.02**	2.01	0.37	1.56	0.45	0.46	0.17	0.31	0.61	0.43
Cross-lingual Fr	-	**16.21**	**22.93**	1.27	2.53	0.24	**2.36**	**2.31**	**2.78**	0.16	0.24	**2.23**	2.36
Cross-lingual Zh	**24.82**	14.41	21.36	-	**3.39**	**2.28**	1.99	1.34	1.33	**0.24**	**0.94**	1.36	**2.47**
Bilingual†*	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
Joint†‡	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

Table 3 Example sentences from the test set of De-En and Ja-En ST. Transcription and Reference are given in the dataset, and others were predicted during decoding. Colored words are referred to in Section 4.6 and the same color corresponds to the same meaning. *Encoder initialized with English ASR.

	De
Transcription	Von der Erde sieht man immer dieselbe Seite des Mondes.
Reference	From the earth you always see the same side of the moon.
Bilingual*	You are always looking from the earth of the moon.
Cross-lingual Fr	From the earth, one always sees the same side of the moon.
Cross-lingual Zh	Of the earth you always look like this side of the moon.
	Ja
Transcription	父は木にはしごを立てかけた。
Reference	Father set the ladders against the tree.
Bilingual*	My father husband home home home.
Cross-lingual Fr	The lady treated that man magnanimously.
Cross-lingual Zh	Father trained to be a cold.

can also see that cross-lingual Fr predicted *sees* correctly, but others were like *looking from* or *look like*.

In Ja-En ST, the sentence from the bilingual baseline is not grammatically correct. On the other hand, the sentence produced by our method is more natural English, although the meaning of it is far from the original one. Cross-lingual Zh is better than cross-lingual Fr in that it correctly predicts *Father*. We can say that our method is effective for both high-resource settings and low-resource settings.

5 Conclusion and Future Work

We proposed end-to-end ST with cross-lingual transfer learning, which is effective for both low-resource and high-resource settings, especially when the knowledge is transferred from a linguistically similar language. In the future, we will explore the joint model considering linguistic similarity. We will also explore which layer of the encoder/decoder at stage 2 has effective information to transfer.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.

References

[1] Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7409–7421, Online, July 2020. Association for Computational Linguistics.

[2] Changhan Wang, Anne Wu, and Juan Pino. CoVoST2 and Massively Multilingual Speech-to-Text Translation. *arXiv preprint arXiv:2007.10310v3*.

[3] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. fairseq S2T: Fast Speech-to-Text Modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*, 2020.

[4] Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 570–577. IEEE, 2019.

[5] Ankur Bapna, Colin Andrew Cherry, Dmitry (Dima) Lepikhin, George Foster, Maxim Krikun, Melvin Johnson, Mia Chen, Naveen Ari, Orhan Firat, Wolfgang Macherey, Yonghui Wu, Yuan Cao, and Zhifeng Chen. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv:1907.05019*, 2019.

[6] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine 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)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, Vol. 53, No. 5, September 2020.

[8] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training 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)*, pp. 58–68, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7909–7913. IEEE, 2020.

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

[11] Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3728–3738, 2020.

[12] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6224–6228, 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[14] 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 *Proc. Interspeech 2019*, pp. 2613–2617, 2019.

[15] Taku Kudo and John Richardson. 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*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

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