ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095811

JOINT TRAINING AND DECODING FOR MULTILINGUAL END-TO-END SIMULTANEOUS SPEECH TRANSLATION

Wuwei Huang[1*] *Renren Jin*[2] *Wen Zhang*[1] *Jian Luan*[1] *Bin Wang*[1] *Deyi Xiong*[2†]

[1]Xiaomi AI Lab, Beijing, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China

ABSTRACT

Recent studies on end-to-end speech translation(ST) have facilitated the exploration of multilingual end-to-end ST and end-to-end simultaneous ST. In this paper, we investigate end-to-end simultaneous speech translation in a one-to-many multilingual setting which is closer to applications in real scenarios. We explore a separate decoder architecture and a unified architecture for joint synchronous training in this scenario. To further explore knowledge transfer across languages, we propose an asynchronous training strategy on the proposed unified decoder architecture. A multi-way aligned multilingual end-to-end ST dataset was curated as a benchmark testbed to evaluate our methods. Experimental results demonstrate the effectiveness of our models on the collected dataset. Our codes and data are available at: `https://github.com/XiaoMi/TED-MMST`.

Index Terms— End-to-End ST, Simultaneous Machine Translation, Multilingual

1. INTRODUCTION

End-to-end speech translation (ST) directly translates speech utterances in source language into texts in target language. Due to its advantages over cascade ST that suffers from error propagation and information loss [1], end-to-end ST has recently attracted growing attention and made substantial progress [2].

Such progress has further promoted the exploration of end-to-end ST in many realistic scenarios. One of them is multilingual scenario, where speech utterances in one or multiple source languages are translated into multiple target languages, e.g., in online lectures or meetings. Inspired by multilingual neural machine translation (NMT), where a single learning system is trained for multiple language pairs [3, 4], multilingual end-to-end ST has been investigated [5, 6]. The nature of multi-task learning and transfer learning across languages substantially benefit multilingual end-to-end ST, especially for low-resource languages.

Yet another real-world scenario for end-to-end ST is simultaneous translation. Simultaneous speech translation begins to generate target words before the entire speech input is received [7, 8], requiring a trade-off between translation quality and latency. In this paper, we investigate both multilinguality and simultaneousness in end-to-end ST, i.e., one-to-many multilingual end-to-end simultaneous ST, which finds its applications in various scenarios, e.g., international conferences, online multilingual conversation, online lectures for students from different countries. These scenarios typically require simultaneously translating speech signals in one source language into multiple target languages. The combination of multilinguality and simultaneousness in end-to-end ST is nontrivial as it is confronted with challenges from both data scarcity and simultaneously decoding among different languages. We investigate two neural architectures for joint multilingual end-to-end simultaneous ST, a separate decoder model and a unified encoder-decoder model in a joint synchronous training. Specifically, the separate decoder model uses different decoders for different target languages, while the unified model shares all decoders across all target languages. To further explore knowledge transfer, we devise joint asynchronous training.

Our contributions can be summarized as follows:

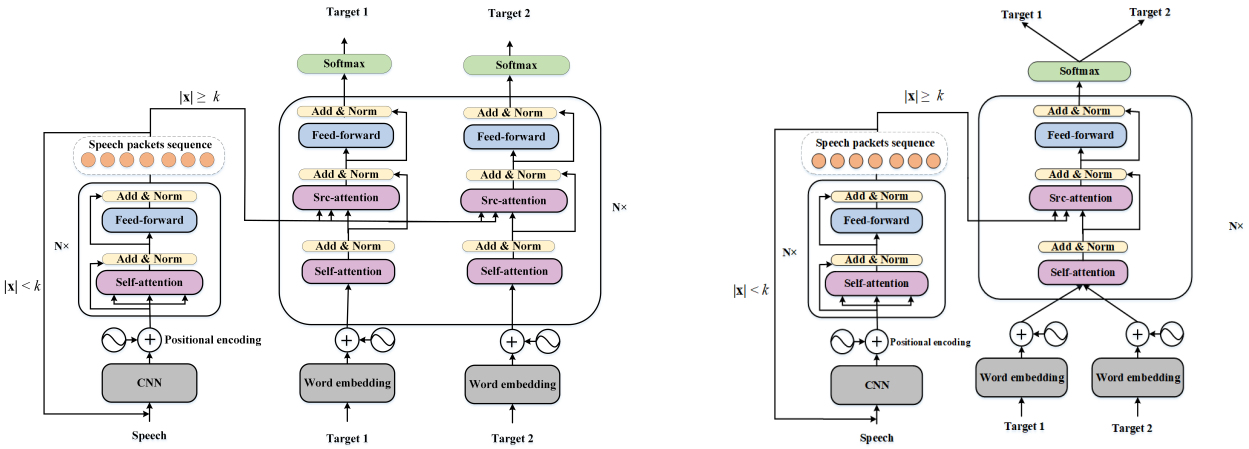
- We present two models in a synchronous training for multilingual end-to-end simultaneous ST.
- In addition to standard synchronous training, we further propose asynchronous training to explore knowledge transfer between different languages.
- We curate a dataset as a benchmark for multilingual end-to-end simultaneous ST.
- Our experiments demonstrate the effectiveness of the proposed framework and training strategy for multilingual end-to-end simultaneous speech translation.

2. RELATED WORK

Since the pioneering works on end-to-end ST [9, 10], a wide range of methods have been proposed. [11] and [12] take advantages of multitask learning to train end-to-end speech translation model by using automatic speech recognition (ASR) or machine translation (MT) as auxiliary tasks. To exploit

[*] The work was done while the first author was a postgraduate student at Tianjin University.

† Corresponding author.



(a) Separate Decoder Model (b) Unified Model

Fig. 1. Diagram of joint multilingual end-to-end simultaneous speech translation.

rich resources in ASR and MT, [13] synthesizes training data for end-to-end speech translation from ASR and MT datasets. [14]and [15] optimize speech input modeling in the encoder by reducing the length of acoustic sequences. Knowledge distillation[16] and curriculum learning [17] have also been explored in end-to-end ST.

To overcome the issue of data scarcity, [6] mixes data of different languages pairs to train a one-to-many multilingual end-to-end ST model. [5] investigates both one-to-many and many-to-many multilingual end-to-end ST.

Studies in simultaneous machine translation greatly inspire the exploration of simultaneous speech translation. [7] adapts methods originally proposed for simultaneous machine translation to end-to-end simultaneous ST. None of the aforementioned previous studies on either multilingual end-to-end ST or end-to-end simultaneous ST investigate the nontrivial conjunction of these two technologies. The most recent work similar to ours is [18], but it trains models neither synchronously nor asynchronously.

3. METHOD

We propose two joint training architectures for multilingual end-to-end simultaneous ST, which vary the degree to which parameters are shared across languages, as shown in Figure 1. Without loss of generality, we take two target languages as a special case to discuss our methods.

3.1. Separate Decoder Model

Our separate decoder model for multilingual end-to-end simultaneous ST shares speech encoder parameters across different languages. The training loss for the j^{th} target language is:

$$\mathcal{L}^j\left(\mathbf{x}, \mathbf{y}^j; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d^j\right) = \sum_{t=1}^{|\mathbf{y}^j|} -\log p\left(\mathbf{y}_t^j | \mathbf{y}_{<t}^j, \mathbf{x}_{<g(t)}; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d^j\right) \quad (1)$$

where $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_d^j$ represent the parameters of the encoder and the j^{th} decoder, respectively. $\mathbf{y}^j$ means the ground-truth translation corresponding to the j^{th} target language, and $\mathbf{y}_{<t}$ is the first t-1 tokens of the j^{th} target translation. Decoders for different languages do not share parameters. $g(t)$ denotes the length of source speech packets read by the model before writing the t^{th} target token. The separate decoder model is similar to [19], which explores interactive decoding between two tasks or two target languages. Our motivation is significantly different from theirs. We aim at cross-lingual knowledge transfer, rather than using the information of another task or another language for decoding.

3.2. Unified Model

The second model for multilingual end-to-end simultaneous ST shares all parameters across different languages in a unified model. To enable the model to distinguish different target languages during decoding, a token identifying the target language is prepended to the target sequence of each training instance. The training loss for the unified model for different languages can be formulated as follows

$$\mathcal{L}\left(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d\right) = \sum_{t=1}^{|\mathbf{y}|} -\log p\left(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}_{<g(t)}; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d\right) \quad (2)$$

where the decoders for different languages share the same parameters $\boldsymbol{\theta}_d$.

3.3. Joint Training

The standard training *(i.e., joint synchronous training)* is used for both the separate decoder model and the unified model. We also propose a novel training strategy: *joint asynchronous training*. Without loss of generality, we take two target languages as an example.

3.3.1. Joint Synchronous Training

To deal with simultaneousness, we adopt wait-k strategy and fixed pre-decision module [7]. Our model first reads k speech packets, each of which contains fixed q speech frames where q is a hyper-parameter in the fixed pre-decision module. The decoder begin to generates one token simultaneously for each language after k speech packet is read. Since the same k value is employed on different target languages, we call this training method *joint synchronous training*. Once a speech packet is accepted, the operation taken by the decoder is formalized as follows:

$$\textbf{Operation} = \begin{cases} \text{continue to read} & |\mathbf{x}| < k \\ \text{output } \mathbf{y}_t^1,\ \mathbf{y}_t^2 & |\mathbf{x}| \geq k \end{cases} \tag{3}$$

where $\boldsymbol{y}_t^1$ and $\boldsymbol{y}_t^2$ represent the t^{th} token of each target language, and $|\mathbf{x}|$ represent the length of speech packet. Our final training objective is:

$$\begin{aligned} &\mathcal{L}(\mathbf{x}, \mathbf{y}^1, \mathbf{y}^2; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d^1, \boldsymbol{\theta}_d^2) = \\ &\quad \mathcal{L}^1\left(\mathbf{x}, \mathbf{y}^1; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d^1\right) + \mathcal{L}^2\left(\mathbf{x}, \mathbf{y}^2; \boldsymbol{\theta}_e, \boldsymbol{\theta}_d^2\right) \end{aligned} \tag{4}$$

$\mathcal{L}^1$ and $\mathcal{L}^2$ are the loss functions for each language, as defined by Eq. 1.

3.3.2. Joint Asynchronous Training

As introduced in the previous subsection, multiple languages use the same k during synchronous training. In actual usage scenarios, each target language holds an appropriate k to balance the translation quality and latency. In many cases, these k values are not identical across languages. We hence propose an asynchronous training strategy.

Once a speech packet is received, the subsequent operation of the decoder can be defined as follows:

$$\textbf{Operation} = \begin{cases} \text{continue to read} & |\mathbf{x}| < k^1, k^2 \\ \text{output } \mathbf{y}_t^1 & k^1 \leq |\mathbf{x}| < k^2 \\ \text{output } \mathbf{y}_t^1,\ \mathbf{y}_t^2 & |\mathbf{x}| \geq k^1, k^2 \end{cases} \tag{5}$$

where k^1 and k^2 are the k values corresponding to the two languages, k^1 is smaller than k^2.

In order clearly describe the process of asynchronous translation, we compare asynchronous training with synchronous training. In the asynchronous schema, we assume the case of k=4 for English-to-Spanish and k=6 for English-to-French, while in synchronous schema, we set k to 6. The decoder in the asynchronous schema starts to generate a Spanish translation with a delay of 4 speech packets, and starts to generate a French translation with a delay of 6 speech packets. In contrast to the decoder in the asynchronous schema, the decoder in the synchronous schema begins to generate Spanish and French translations simultaneously with a delay of 6 speech packets.

Domain	Hours	Sentences			Direction
		train	dev	test	
TED	350	171K	1.5K	2.7K	En→Fr, Es

Table 1. Statistics of the curated data for multilingual end-to-end simultaneous speech translation.

4. DATA CURATION

Most previous simultaneous speech translation research is based on the MuST-C [20] dataset, which comes from TED talks. Unfortunately MuST-C does not include data where source speech signal is translated into multiple target languages. Thus, we curate a multilingual simultaneous speech translation dataset from TED website.

TED website offers English talks on various topics, as well as translations into different languages. In this paper, we use French and Spanish as target languages for simultaneous training. To achieve this, we crawled approximately 2.5K speech audios and the corresponding English transcripts, French and Spanish translated texts from TED. After collecting the raw data, a two-step alignment should be performed to meet the training requirements. First, we preformed source sentence segmentation based on hard punctuation, using Gentle[1] to achieve audio-to-transcript alignments. Second, we created alignments that are unique to each En-XX language pair by Laser[2] and finally obtained the intersection of the alignmnets across target texts. We split the final 175K sentences into 171K sentences as the training set, 1.5K as the dev set and 2.7K as the test set. The details of the dataset are shown in Table 1. We also conducted offline experiments and the BLEU score for En-Es and En-Fr are 28.56 and 24.46, respectively.

5. EXPERIMENTS

We conducted experiments to evaluate the effectiveness of the proposed models on the curated dataset.

5.1. Evaluation Metrics

Typically, simultaneous speech translation is evaluated with both translation quality and latency. The former was measured with detokenized BLEU [21] in our experiments. In terms of latency, average latency (AL) [22], average proportion (AP) [23] and Differentiable Average Lagging (DAL) [7] are commonly used. Under the simultaneous mechanism adopted by this paper, the latency is positively correlated with k. In order to represent the results of asynchronous experiments more clearly, we used k as evaluation metric for latency throughout this paper.

[1] https://github.com/lowerquality/gentle

[2] https://github.com/facebookresearch/LASER

Tasks	Models	k=3	k=4	k=5	k=6
En→Es	Bilingual	7.90	11.13	12.82	15.80
	Separate Decoder	9.47	12.77	15.62	17.39
	Unified	11.09	12.62	15.89	17.55
En→Fr	Bilingual	10.09	13.71	15.59	16.87
	Separate Decoder	11.57	14.20	16.66	17.67
	Unified	13.39	14.69	17.35	18.19

Table 2. BLEU scores on En-Es and En-Fr corresponding to k in SimulEval with 3, 4, 5, 6, respectively.

Training Strategy	k_{Es}	k_{Fr}	En→Es	En→Fr
Synchronous	4	4	12.62	14.69
	6	6	17.55	18.19
Asynchronous	6	4	17.77	15.74
	4	6	14.40	18.61

Table 3. BLEU scores of asynchronous translation on En-Es and En-Fr.

5.2. Settings

We built our models based on the Fairseq[3] toolkit. We extracted 80-dimensional Fbank features from audio files. We used a shared vocabulary with a size of $8K$ tokens and fixed the vocabulary for all experiments. The dimension of the attention was set to 256. We used 12-layer encoder and the speech encoder are initialized based on a pre-trained ASR task. The hyper-parameter q introduced in Section 3.3.1 wet set to 7. The number of decoder layers was set to 6. We trained 70 epochs for each model and used the Adam optimizer. All models were run on four NVIDIA RTX A6000 GPUs. After training, we test our model on SimulEval[4] with greedy decoding strategy.

6. RESULTS AND ANALYSES

6.1. Performance of the Separate Decoder Model and Unified Model

Table 2 shows the comparison results of the proposed separate decoder model and unified model against the standard bilingual translation model with varying k. It is clear that both proposed models for multilingual end-to-end simultaneous ST are able to significantly improve translation quality. Particularly, the separate decoder model achieves an average improvement of 1.90 BLEU points on English-to-Spanish ST, and 0.96 points on English-to-French. In the case of the unified model, the gains on English-to-Spanish ST and English-to-French ST are 2.38 and 1.84 BLEU, respectively. In terms of improvements, the unified model is superior to the separate decoder model even with fewer parameters than the latter, indicating full parameter sharing enhances transfer learning.

[3]https://github.com/pytorch/fairseq

[4]https://github.com/facebookresearch/SimulEval

6.2. Asynchronous Translation

When the asynchronous training schema is taken, we adopt the unified model. Two settings are conducted: one is setting the k to 6 for English-to-Spanish direction and setting k to 4 for English-to-French while the other is setting the k to 4 for English-to-Spanish direction and setting the k to 6 for English-to-French.

We compare asynchronous strategy with synchronous strategy and the results are shown in Table 3. The asynchronous training improves translation quality at each k. When setting the k to 6 for English-to-Spanish direction and 4 for English-to-Spanish direction, the asynchronous approach gains an improvement of 1.05 for English-to-French translation, and 0.22 BLEU for English-to-Spanish translation over the synchronous training. The other asynchronous training setting also results in improvement of 1.78 for English-to-Spanish translation, and 0.42 BLEU for English to French translation over the synchronous training. These proves the effectiveness of the joint asynchronous training strategy.

It is generally acknowledged that in the prefix-to-prefix scenario, as we train the model to predict using source prefixes, the trained streaming model is often able to perceive preceding context which we call $anticipation$ [22]. We believe that with asynchronous training, the target language with latency k^1 in the model can increase its $anticipation$ ability due to the existence of the target language with k^2. e.g., in joint synchronous training scenario or bilingual scenario, the target language with fixed k has limited $anticipation$ ability. While in joint asynchronous training scenario, the target language with a larger k plays as a transitional role, so that the target language with a smaller k has an access to get the $anticipation$ ability of the target with a larger k.

7. CONCLUSIONS

In this work, we have investigated joint training and decoding for multilingual end-to-end simultaneous speech translation and curated a dataset to train and evaluate the proposed methods. Experiments demonstrate that the proposed separate decoder model and unified model significantly improve translation quality of simultaneous end-to-end speech translation on a synchronous training setting. Finally, we find the joint asynchronous training method further improves translation quality.

8. ACKNOWLEDGMENTS

The present research was supported by the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004). We would like to thank the anonymous reviewers for their insightful comments.

9. REFERENCES

[1] Matthias Sperber and Matthias Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," in *Proc. of ACL*, 2020.

[2] Wuwei Huang, Dexin Wang, and Deyi Xiong, "Adast: Dynamically adapting encoder states in the decoder for end-to-end speech-to-text translation," in *Proc. of ACL Findings*, 2021.

[3] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proc. of NAACL*, 2016.

[4] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlícek, Tanja Schultz, and Hervé Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proc. of ICASSP*, 2014.

[5] Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe, "Multilingual end-to-end speech translation," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, 2019.

[6] Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi, "One-to-many multilingual end-to-end speech translation," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, 2019.

[7] Xutai Ma, Juan Miguel Pino, and Philipp Koehn, "Simulmt to simulst: Adapting simultaneous text translation to end-to-end simultaneous speech translation," in *Proc. of AACL*, 2020.

[8] Xutai Ma, Yongqiang Wang, Mohammad Javad Dousti, Philipp Koehn, and Juan Miguel Pino, "Streaming simultaneous speech translation with augmented memory transformer," in *Proc. of ICASSP*, 2021.

[9] Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *CoRR*, vol. abs/1612.01744, 2016.

[10] Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li, "Listen, understand and translate": Triple supervision decouples end-to-end speech-to-text translation," *CoRR*, vol. abs/2009.09704, 2020.

[11] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proc. of INTERSPEECH*, 2017.

[12] Alexandre Berard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proc. of ICASSP*, 2018.

[13] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proc. of ICASSP*, 2019.

[14] Elizabeth Salesky, Matthias Sperber, and Alan W. Black, "Exploring phoneme-level speech representations for end-to-end speech translation," in *Proc. of ACL*, 2019.

[15] Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich, "Adaptive feature selection for end-to-end speech translation," in *Proc. of EMNLP Findings*, 2020.

[16] Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong, "End-to-end speech translation with knowledge distillation," in *Proc. of INTERSPEECH*, 2019.

[17] Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang, "Curriculum pre-training for end-to-end speech translation," in *Proc. of ACL*, 2020.

[18] Jian Xue, Peidong Wang, Jinyu Li, Matt Post, and Yashesh Gaur, "Large-scale streaming end-to-end speech translation with neural transducers," in *Proc. of INTERSPEECH*, 2022.

[19] Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier, "Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation," in *Proc. of COLING*, 2020.

[20] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. of NAACL*, 2019.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002.

[22] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang, "STACL: simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework," in *Proc. of ACL*, 2019.

[23] Kyunghyun Cho and Masha Esipova, "Can neural machine translation do simultaneous translation?," *CoRR*, 2016.