Interspeech 2022
18-22 September 2022, Incheon, Korea

Cross-Modal Decision Regularization for Simultaneous Speech Translation

Mohd Abbas Zaidi⋆, *Beomseok Lee*⋆, *Sangha Kim*, *Chanwoo Kim*

Samsung Research, Seoul, South Korea

{abbas.zaidi, bsgunn.lee, sangha01.kim, chanw.com}@samsung.com

Abstract

Simultaneous translation systems start producing the output while processing the partial source sentence in the incoming input stream. These systems need to decide when to *read* more input and when to *write* the output. The decisions taken by the model depend on the structure of source/target language and the information contained in the partial input sequence. Hence, *read/write* decision policy remains the same across different input modalities, i.e., speech and text. This motivates us to leverage the text transcripts corresponding to the speech input for improving simultaneous speech-to-text translation (SimulST). We propose Cross-Modal Decision Regularization (CMDR) to improve the decision policy of SimulST systems by using the simultaneous text-to-text translation (SimulMT) task. We also extend several techniques from the offline speech translation domain to explore the role of SimulMT task in improving SimulST performance. Overall, we achieve 34.66% / 4.5 BLEU improvement over the baseline model across different latency regimes for the MuST-C English-German (EnDe) SimulST task.

Index Terms: speech translation, simultaneous translation, decision policy.

1. Introduction

1.1. Simultaneous Translation

Simultaneous translation systems find huge applications in real life scenarios such as live subtitle generation and real-time interpretation. To provide translation in tandem with the streaming input, these systems alternate between *read/write* decisions, i.e., *reading* the source sequence and *writing* the target tokens. Initial approaches for such systems, such as wait-k [1] use a fixed policy, where the *read/write* schedule is pre-decided. Recent works [2, 3] use monotonic attention [4, 5] to learn a flexible decision policy. Monotonic attention provides a closed form expression for the expected input output alignment to train the discrete *read/write* decisions in expectation. Monotonic multihead attention (MMA) [3] replaces the soft attention in the transformer [6] model with monotonic attention outperforming the models with fixed decision policy.

1.2. Offline Speech Translation

The offline speech translation (ST) task has traditionally suffered from data scarcity issues. Hence, various approaches use pretraining [7], multitask learning [8, 9], meta-learning [10] and knowledge distillation (KD) [11] to leverage the high resource machine translation (MT) and automatic speech recognition (ASR) tasks for improving ST performance. A recent work [9] improves information sharing between offline MT and ST task by using online KD and cross-attentive regularization (CAR). These approaches have demonstrated that the MT task can be crucial to improve the performance of ST systems.

⋆Equal contribution.

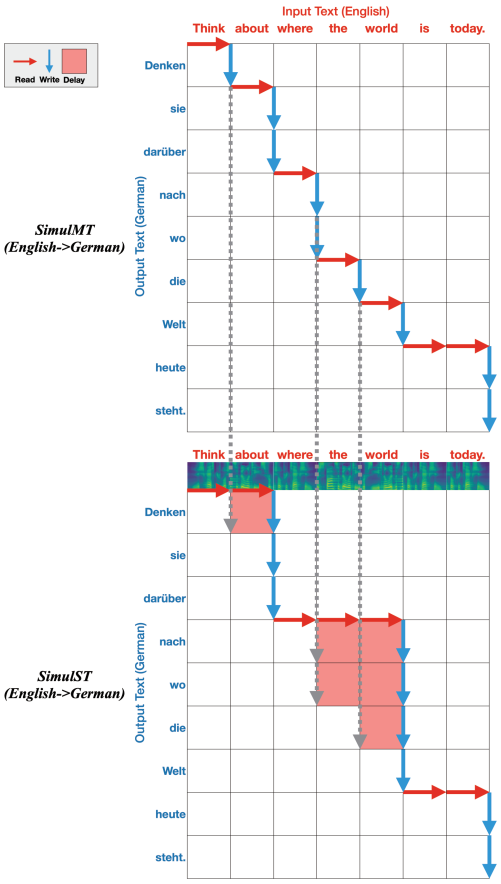

Figure 1: *An example of using SimulMT decision to improve SimulST for English-German translation. SimulMT can prevent the SimulST model from incurring extra delays.*

1.3. Simultaneous Speech Translation

Simultaneous speech translation (SimulST) systems perform realtime speech-to-text translation. A recent work [12] brings together the advances in transformer [6] based offline ST systems and montonic multihead attention (MMA) based simultaneous translation.

For MMA-based SimulST model, *read/write* decisions are guided by the monotonic attention energies learned during training. In the absence of direct supervision for the decision policy, the MMA model learns these decisions by balancing the trade-off between output accuracy and latency. These *read/write* decisions depend on the amount of information contained in the source sequence, and the word orders of the source/target languages. Hence, the decision policy for speech and text inputs remains the same. Moreover, due to the relative complexity associated with speech inputs, it is easier to learn the *read/write*

10.21437/Interspeech.2022-10617

decisions from the text data. Hence, SimulMT decisions can potentially serve as a reference to improve the SimulST decision policy as shown in Figure 1.

We propose Cross-Modal Decision Regularization (CMDR) which utilizes the monotonic attention energies of the SimulMT model to guide the decision policy of the SimulST model implicitly. We also extend several techniques from the offline ST domain, such as multitask learning, online KD and CAR to the SimulST task. Experiments on the MuST-C EnDe dataset show that the proposed CMDR, along with other approaches improves the performance of MMA-based SimulST systems significantly.

2. MODEL

We use the MMA model described in [12] as our baseline SimulST model. It processes partial input speech and partial target text to produce the next target token. Given partial source and target ($x_{\leq j} \in \mathbf{x}, y_{<i} \in \mathbf{y}$), the next target token y_i is generated as follows:

$$h_j = \mathcal{E}(x_{\leq j}) \tag{1}$$

$$s_i = \mathcal{D}(y_{<i}, \mathcal{MA}(s_{<i}, h_{\leq j})) \tag{2}$$

$$y_i = Output(s_i) \tag{3}$$

where $\mathcal{E}(.)$ and $\mathcal{D}(.)$ represent the encoder and decoder blocks and $\mathcal{MA}$ refers to the monotonic multihead attention energy function. As seen in Figure 1, the MMA model alternates between *read* and *write* decisions at test time. It uses the monotonic attention energies to make *read/write* decisions as follows:

$$e_{i,j} = \mathcal{MA}(s_{i-1}, h_j) \tag{4}$$

$$p_{i,j} = Sigmoid(e_{i,j}) \tag{5}$$

$$z_{i,j} \sim Bernoulli(p_{i,j}) \tag{6}$$

When $z_{i,j} = 1$ (*write* decision), the model sets $t_i = j$ and computes the decoder output using $h_{\leq j}$, where t_i refers to the number of encoder states required to produce the i_{th} decoder output. If $z_{i,j} = 0$, the model needs to read further. It computes h_{j+1} and repeats Eq. 4 to 6.

As mentioned earlier, the proposed CMDR loss aims to improve the SimulST decision policy using the monotonic energy activation of the SimulMT model. CMDR computes the similarity between the monotonic energies of speech and text input corresponding to each training example. However, it cannot be computed directly since the attention energies corresponding to speech ($\mathbf{A}^s$) and text ($\mathbf{A}^t$) have different sizes due to different input lengths. Similar to [9], we use self-attention and cross-attention operations with respect to $\mathbf{A}^t$ to obtain attention representations $\mathbf{A}^{t\to t}$ and $\mathbf{A}^{s\to t}$ which have the same size. Finding the $\mathcal{L}2$ distance between these reconstructed representations provides the required cross-modal similarity metric for each example.

During joint training of SimulST and SimulMT, each training example consists of input speech and corresponding transcript in the source language and output text in the target language. Let K and L denote the length of speech and text representations at the output of the encoder. The monotonic attention energy matrices for speech and text for the h_{th} head are defined as follows:

$$\mathbf{A}^s_h = (at^s_{h,1}, \cdot\cdot, at^s_{h,K}), \mathbf{A}^t_h = (at^t_{h,1}, \cdot\cdot, at^t_{h,L})$$

where $at_j = e_{(i,:)}$ refers to the monotonic attention corresponding to the j_{th} encoder output token, aggregated across all decoder indices. Attention energies from H different heads are stacked as follows:

$$\mathbf{A}^s = [\mathbf{A}^s_1, \ldots, \mathbf{A}^s_H] = (a^s_1, a^s_2, \ldots, a^s_{K\times H}) \tag{7}$$

$$\mathbf{A}^t = [\mathbf{A}^t_1, \ldots, \mathbf{A}^t_H] = (a^t_1, a^t_2, \ldots, a^t_{L\times H}) \tag{8}$$

Next, similarity matrix $\mathbf{S}$ is used to obtain $\mathbf{A}^{s\to t}$ via cross-attention between $\mathbf{A}^s$ and $\mathbf{A}^t$. Similarly, $\mathbf{A}^{t\to t}$ is obtained from $\mathbf{A}^t$ by using self-attention.

$$s_{i,j} = \frac{a^s_i \cdot a^t_j}{\|a^s_i\|_2 \left\|a^t_j\right\|_2} \tag{9}$$

$$\mathbf{A}^{s\to t} = \mathbf{A}^s \cdot \text{softmax}(\mathbf{S}) \tag{10}$$

For the d_{th} decoder layer, the CMDR loss is computed as follows:

$$\mathcal{L}^d_{CMDR}(\theta_s) = \sum \frac{1}{LH} \left\|\mathbf{A}^{s\to t} - sg[\mathbf{A}^{t\to t}]\right\|_2 \tag{11}$$

where sg (stop gradient operator) allows the model to use text attention as a reference for the speech attention. Finally, the CMDR loss is computed by averaging across M decoder layers.

$$\mathcal{L}_{CMDR}(\theta_s) = \sum_{d=1}^{M} \frac{1}{M} \mathcal{L}^d_{CMDR}(\theta_s) \tag{12}$$

In addition to the proposed CMDR approach, this work also extends several existing techniques from the offline ST domain. Firstly, it employs multitask learning by training SimulMT model along with SimulST. It also extends online KD and CAR [9] losses to SimulST. Finally, similar to other MMA-based translation systems, it uses differentiable average lagging (DAL) [13] loss to control the latency of simultaneous translation models. The overall loss $\mathcal{L}(\theta_s, \theta_t)$ is defined as follows:

$$\begin{aligned}\mathcal{L}(\theta_s, \theta_t) = (1-\alpha)\mathcal{L}_{ST-NLL}(\theta_s) + \alpha\mathcal{L}_{KD}(\theta_s, \theta_t) \\ + \beta\mathcal{L}_{CAR}(\theta_s) + \gamma\mathcal{L}_{MT-NLL}(\theta_t) \\ + \delta\mathcal{L}_{CMDR}(\theta_s, \theta_t) + \lambda\mathcal{L}_{DAL}(\theta_s, \theta_t)\end{aligned} \tag{13}$$

It combines the negative-log likelihood loss for both speech (ST) and text (MT) with KD, CAR, CMDR, and DAL loss. (θ_s, θ_t: speech/text model parameters)

3. EXPERIMENTAL SETTINGS

3.1. Dataset

For SimulST, MuST-C [14] English-German (En-De) dataset is used for training with tst-COMMON as the test set. For SimulMT, WMT 14 [15] and MuST-C En-De serve as the training data. Table 1 provides the dataset statistics. Data preprocessing details are the same as [9].

Task	# Hours	# Sentences		
		Train	Dev	Test
MuST-C	408	225k	1,423	2,641
WMT 14	-	2.57M	26k	3003

Table 1: *Dataset Statistics (# - Number of)*

Hyper-parameter	Value
speech conv layers	2
speech conv stride	(2,2)
shared encoder layers	6
encoder embed dim	512
encoder ffn embed dim	2048
encoder attention heads	8
decoder embed dim	512
decoder ffn embed dim	2048
decoder attention heads	8
dropout	0.1
optimizer	adam
adam-β	(0.9, 0.999)
clip-norm	10.0
lr scheduler	inverse sqrt
learning rate	0.002
warmup-updates	20000
label-smoothing	0.1
max text tokens per batch	5000
max speech frames per batch	5000

Table 2: *List of Hyperparameters*

3.2. Data Augmentation

Equal amounts of synthetic speech and text data is generated for MuST-C EnDe dataset. Augmented speech is paired with original text and vice versa. Synthetic speech is generated by varying the SoX effects similar to [10], while the augmented target text is generated by translating [16] the source transcripts using the WMT 19 winner offline MT En-De model [17].

3.3. Pretraining and Weight Sharing

We use MMA-based transformer as our base model. The transformer decoder of the offline ST model [9] is replaced with a monotonic decoder [12]. The base model consists of a speech encoder (12 layers), text encoder (6 layers), and a joint monotonic decoder (6 layers) shared between the SimulST and SimulMT models. Similar to [9], the top 6 layers of the speech encoder are tied to the text encoder. The speech encoder is initialized using pretrained ASR encoder[1]. MT encoder and joint decoder are initialized using an offline MT [2] model.

3.4. Training Details

All the models are implemented using the Fairseq [18] toolkit. In order to compute the CMDR loss, parallel ST and MT data is required. The speech transcripts in the MuST-C EnDe dataset are utilized as the parallel SimulMT input required to compute the cross-modal similarity losses.

All the models are trained using 8xA100 GPUs with an update frequency of 4. The training schedule is the same as [12]. The model is first trained for 150 epochs (110 hours / 4.5 days) without the differentiable average lagging (DAL) latency loss by setting $\lambda = 0$ in Eq. 13. These models are referred to as λ_0 models. Finally, the models are finetuned further for 50 epochs (40 hours) after adding the DAL loss. The training with the latency loss is carried out using three different values of $\lambda \in \{0.01, 0.05, 0.1\}$. During inference, the step sizes are varied ($\{120, 200, 280, 360, 440, 520\}$ in ms) to obtain the model performance in different latency regimes. Step-size / speech-segment size refers to the duration of speech consumed corresponding to each *read* decision. The weights for various losses in Eq. 13 are as follows: $\alpha = 0.2$ for online KD, $\beta = 0.02$ for CAR and $\gamma = 0.5$ for MT-NLL. $\delta = 0.01$ is the weight for the proposed CMDR loss. It was obtained using standard grid-search from 0.05 to 0.5 with a step-size of 0.05. Detailed hyperparameter settings required to replicate the results can be found in the Table 2.

[1]ASR Model: `https://dl.fbaipublicfiles.com/fairseq/s2t/mustc_joint_asr_transformer_m.pt`

[2]MT Model: `https://dl.fbaipublicfiles.com/joint_speech_text_4_s2t/must_c/en_de/checkpoint_mt.pt`

4. RESULTS

As mentioned in Section 2 and 3, this work extends data augmentation, multitask learning, online KD and CAR based techniques from offline ST domain to the SimulST task. It also proposes a new CMDR loss for SimulST systems. Table 3 reports the performance and improvements for the λ_0 models trained using different methods. In order to plot the latency-quality curves [1], multiple models are trained with different λ values. Further, during inference, step-sizes are varied to obtain model performance in different latency regimes. Case-sensitive detokenized BLEU [19] is used to measure the quality while average lagging (AL) [12] is used as the latency metric.

No	Methods	Model Name	λ_0 models	
			BLEU	Δ
1	Baseline	*MMA*	17.23	-
2	1 + Data Augmentation	*MMA-Aug*	19.81	2.58
3	1 + Multitask Learning	-	18.90	1.67
4	3 + Online KD	-	19.85	0.95
5	4 + CAR	*MMA-MT*	20.10	0.25
6	5 + Data Augmentation	*MMA-Aug-MT*	21.49	1.39
This Work				
7	5 + **CMDR**	-	**20.67**	**0.57**
8	6 + **CMDR**	*MMA-CMDR*	**22.35**	**0.86**

Table 3: *Performance of various approaches: λ_0 models*

4.1. Existing Approaches

4.1.1. Data Augmentation

As mentioned in Section 3, several data augmentation techniques are added to the MMA baseline. This model is referred to as *MMA-Aug*. Data augmentation provides significant performance gains both for the baseline and our proposed approach. For λ_0 models, it improves the BLEU score by 2.58 points over *MMA*. Figure 2 provides the latency-quality tradeoff comparison of *MMA-Aug* versus the *MMA* baseline.

4.1.2. Multitask Learning

Multitask learning refers to simply training the SimulST and SimulMT model together with shared parameters. As seen in Table 3, multitask learning boosts the performance of the SimulST task by 1.67 BLEU scores for the λ_0 model.

4.1.3. Online KD & CAR

Similar to the offline domain, online KD and CAR improve the performance of the SimulST model as well. For the λ_0 model, online KD and CAR provide an improvement of 0.95 and 0.25 BLEU scores over the multitask learning baseline.

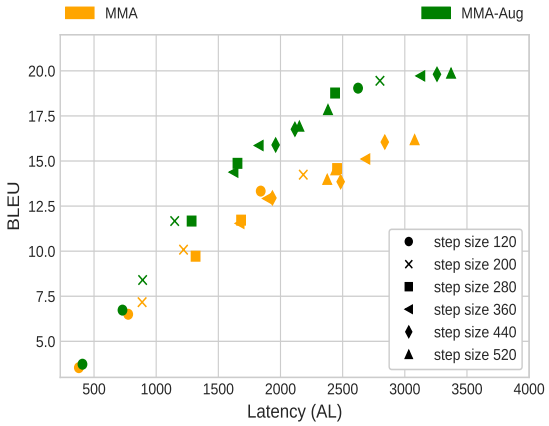

Figure 2: *Effect of Data Augmentation.*

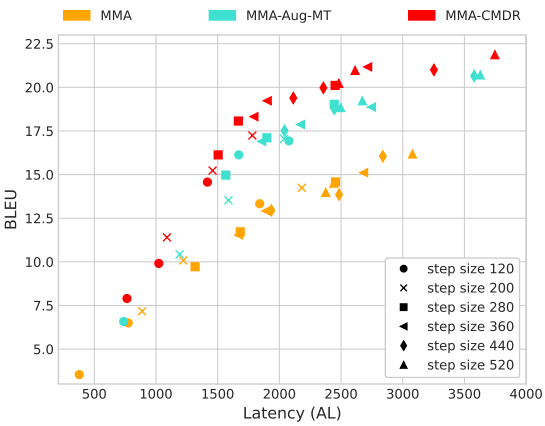

Figure 4: *Effect of the proposed CMDR loss.*

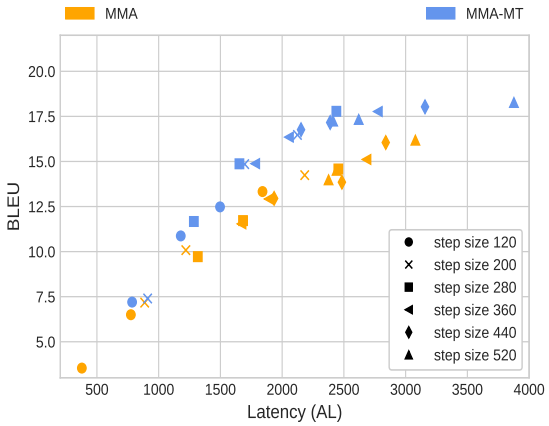

Figure 3: *Effect of existing approaches related to the SimulMT task (Multitask learning, online KD & CAR)*

All the existing techniques related to the auxillary SimulMT task such as multitask learning, online KD, and CAR are grouped together, and this model is referred to as *MMA-MT*. Combined together, these approaches improve the λ_0 model performance by 2.87 BLEU score (Row 5 vs. Row 1 in Table 3). Figure 3 provides the latency-quality curves for *MMA-MT* against the baseline *MMA* model. It provides consistent improvements as compared to the baseline across different latency regimes. Next, another model (*MMA-Aug-MT*) is trained by adding data augmentation techniques to *MMA-MT*. It serves as a baseline to quantify the improvements obtained using the proposed CMDR approach.

4.2. Proposed Approach: CMDR

As discussed in the previous sections, CMDR loss is designed to improve the *read/write* decisions for the challenging SimulST task using the relatively easier SimulMT task. For λ_0 models, CMDR provides a BLEU score improvement of **0.57** over the *MMA-MT* model, and **0.86** over *MMA-Aug-MT*. The final model, *MMA-CMDR* (Row 8 in Table 3) is trained by using all the described approaches (Eq. 13). Figure 4 provides the improvements achieved with respect to the latency-quality tradeoff using CMDR. It provides improvements in the range of $0.7 \sim 1.2$ BLEU scores consistently across all latency regimes. It is interesting to note that the improvements obtained through decision/attention regularization are much higher as compared to regularizing the input speech encoding through CAR.

Validation and Overall results: The improvements obtained from CMDR are statistically significant at 99% confidence with a p value of 0.002 (paired t-test). In addition to the latency-quality curves, we also calculate the overall performance improvement obtained using all the approaches. We choose ten different points with similar latency values and compute the absolute and relative improvements obtained by *MMA-CMDR* over *MMA*. Table 4 reports the exact values used for this comparison. Each of these points is chosen such that the latency for the *MMA-CMDR* model is less than that of *MMA*. Averaging these values, we obtain aggregated improvement of 4.5 BLEU score or 34.66%.

MMA		MMA-CMDR		BLEU	BLEU
BLEU	AL	BLEU	AL	Δ	$\%\Delta$
6.5	775	7.9	765	1.40	21.53
10.08	1220	11.40	1089	1.32	13.09
11.72	1683	16.13	1504	4.41	37.63
13.33	1841	17.24	1781	3.91	29.33
12.92	1891	18.32	1794	5.4	41.79
12.95	1935	19.23	1902	6.28	48.94
14.24	2183	19.39	2113	5.15	36.17
13.98	2376	19.97	2357	5.99	42.84
13.85	2484	20.23	2482	6.38	46.06
16.18	3079	20.97	2614	4.70	29.61

Table 4: *Aggregated performance improvements*

5. CONCLUSIONS

In this work, we leverage the SimulMT task to improve the performance of SimulST system. Various techniques from the offline ST domain, such as online KD and CAR are found to be beneficial for the SimulST task. To improve the performance further, we also propose **Cross-Modal Decision Regularization**. It improves the *read/write* decision policy for SimulST by using the monotonic attention energies of the SimulMT model. This work improves the performance of MMA-based SimulST by 35% or 4.5 BLEU points across different latency regimes.

6. References

[1] M. Ma, L. Huang, H. Xiong, R. Zheng, K. Liu, B. Zheng, C. Zhang, Z. He, H. Liu, X. Li *et al.*, "Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework," *arXiv preprint arXiv:1810.08398*, 2018.

[2] N. Arivazhagan, C. Cherry, W. Macherey, C.-C. Chiu, S. Yavuz, R. Pang, W. Li, and C. Raffel, "Monotonic infinite lookback attention for simultaneous machine translation," *arXiv preprint arXiv:1906.05218*, 2019.

[3] X. Ma, J. Pino, J. Cross, L. Puzon, and J. Gu, "Monotonic multihead attention," *arXiv preprint arXiv:1909.12406*, 2019.

[4] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2837–2846.

[5] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," *arXiv preprint arXiv:1712.05382*, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[7] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

[8] S. Indurthi, M. A. Zaidi, N. Kumar Lakumarapu, B. Lee, H. Han, S. Ahn, S. Kim, C. Kim, and I. Hwang, "Task aware multi-task learning for speech to text tasks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7723–7727.

[9] Y. Tang, J. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," *arXiv preprint arXiv:2107.05782*, 2021.

[10] H. J. Han, M. A. Zaidi, S. R. Indurthi, N. K. Lakumarapu, B. Lee, and S. Kim, "End-to-end simultaneous translation system for iwslt2020 using modality agnostic meta-learning," in *Proceedings of the 17th International Conference on Spoken Language Translation*, 2020, pp. 62–68.

[11] Y. Liu, H. Xiong, Z. He, J. Zhang, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[12] X. Ma, J. Pino, and P. Koehn, "Simulmt to simulst: Adapting simultaneous text translation to end-to-end simultaneous speech translation," *arXiv preprint arXiv:2011.02048*, 2020.

[13] C. Cherry and G. Foster, "Thinking slow about latency evaluation for simultaneous machine translation," *arXiv preprint arXiv:1906.00048*, 2019.

[14] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 2012–2017.

[15] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand *et al.*, "Findings of the 2014 workshop on statistical machine translation," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 12–58.

[16] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," *arXiv preprint arXiv:1808.09381*, 2018.

[17] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook fair's wmt19 news translation task submission," *arXiv preprint arXiv:1907.06616*, 2019.

[18] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[19] M. Post, "A call for clarity in reporting bleu scores," *arXiv preprint arXiv:1804.08771*, 2018.