The AISP-SJTU Simultaneous Translation System for IWSLT 2022

Qinpei Zhu[1] **Renshou Wu**[1] **Guangfeng Liu**[1] **Xinyu Zhu**[1] **Xingyu Chen**[2]
Yang Zhou[1] **Qingliang Miao**[1] **Rui Wang**[2] **Kai Yu**[1,2]
[1]AI Speech Co., Ltd., Suzhou, China
[2]Shanghai Jiao Tong University, Shanghai, China

Abstract

This paper describes AISP-SJTU's submissions for the IWSLT 2022 Simultaneous Translation task. We participate in the text-to-text and speech-to-text simultaneous translation from English to Mandarin Chinese. The training of the CAAT is improved by training across multiple values of right context window size, which achieves good online performance without setting a prior right context window size for training. For speech-to-text task, the best model we submitted achieves 25.87, 26.21, 26.45 BLEU in low, medium and high regimes on tst-COMMON, corresponding to 27.94, 28.31, 28.43 BLEU in text-to-text task.

1 Introduction

This paper describes the systems submitted by AI Speech Co., Ltd. (AISP) and Shanghai Jiaotong University (SJTU) for IWSLT 2022 Simultaneous Translation task. Two speech translation systems including cascade and end-to-end (E2E) for the Simultaneous Speech Translation track, and a simultaneous neural machine translation (MT) system for the text-to-text Simultaneous Translation track. The systems are focused on English to Mandarin Chinese language pair.

For simultaneous speech translation, recent work tends to fall into two categories, cascaded systems and E2E systems. And the cascaded system often outperforms the fully E2E approach. Only one work (Ansari et al., 2020; Anastasopoulos et al., 2021) shows that the E2E model can achieve better results than the cascaded model. In their work they introduce pre-training (Stoian et al., 2020; Dong et al., 2021; Wang et al., 2020b) and data augmentation techniques (Pino et al., 2020; Xu et al., 2021) to E2E models. Therefore, in this paper, we hope to optimize the speech translation model from two aspects. First, we aim to build a robust cascade model and learn best practices from WMT evaluation activities (Wu et al., 2020; Meng et al., 2020; Zeng et al., 2021), such as back translation (Sennrich et al., 2015; Edunov et al., 2018; Lample et al., 2017). Second, we explore various self-supervised learning methods and introduce as much semi-supervised data as possible towards finding the best practice of training cascaded speech-to-text (S2T) models. In our settings, ASR data, MT data, and monolingual text data are all considered in a progressively training framework. We only trained one E2E model, and its BLEU is 22.49 with 1272 AL. Due to the huge difference in the scale of training data from the cascaded model, E2E performance is far lower than that of the latter. The cascaded S2T final performance on the MuST-C V2 test set is 25.87, 26.21, 26.45 BLEU with low, medium and high regimes.

In addition, we also participate in the simultaneous text-to-text (T2T) task. Our system is based on an efficient wait-k model (Elbayad et al., 2020) and CAAT model (Liu et al., 2021b). We investigate large-scale knowledge distillation (Kim and Rush, 2016; Freitag et al., 2017) and back translation methods. Specially, we develop a multi-path training strategy, which enables a unified model serving different wait-k paths. All MT models are based on transformer (Vaswani et al., 2017). The organizers use the output of a streaming ASR system as input to the text-to-text system, and the results will be shown in the overview paper (Anastasopoulos et al., 2022).

The rest of this paper is organized as follows. Section 2 describes the details of the data preprocessing and augmentation. Section 3 describes the models used in our system and introduces details of the model structure and techniques used in training and inference. We present experimental results in Section 4 and related works in Section 5. Finally, the conclusion is given in Section 6.

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 208 - 215

Language	Corpus	Sentences
EN→ZH	WMT2019	20.1M
EN→ZH	WMT2020	20.7M
EN→ZH	WMT2021	42.3M
EN→ZH	OpenSubtitles2018	9.969M
EN→ZH	MuST-C	0.359M

Table 1: Statistics of text parallel datasets.

2 Data Preprocessing and Augmentation

2.1 Data Preprocessing

En-Zh Text Corpora We use English-Chinese (EN-ZH) parallel sentences from WMT2019, WMT2020, WMT2021, OpenSubtitles2018 and MuST-C for training. The statistics of the parallel data is shown in Table 1. Additionally, we select 15% of the Chinese monolingual corpora from News Crawl, News Commentary and Common Crawl for data augmentation. For EN-ZH language pairs, the filtering rules are as follows:

* Filter out sentences that contain long words over 40 characters or over 120 words.
* The word ratio between the source word and the target word must not exceed 1:3 or 3:1.
* Filter out the sentences that have invalid Unicode characters or HTML tags.
* Filter out the duplicated sentence pairs.

Finally, we filter the real and pseudo parallel corpora through a semantic matching model which is trained using limited data. The statistics of the text training data is shown in Table 2.

As for text preprocessing, we apply Moses tokenizer and SentencePiece with 32,000 merge operations on each side.

En-Zh Speech Corpora The speech datasets used in our systems are shown in Table 3, where MuST-C is speech-translation specific (speech, transcription and translation included), and Europarl, CoVoST2, LibriSpeech, TED-LIUM3 and VoxPopuli are speech-recognition specific (only speech and transcription). Kaldi (Ravanelli et al., 2019) is used to extract 80 dimensional log-mel filter bank features, which are computed with a 25ms window size and a 10 ms window shift, and specAugment (Park et al., 2019) are performed during training phase.

	EN→ZH
Bilingual Data	67.4M
Source Mono Data	200.5M
Target Mono Data	405.2M

Table 2: Statistics of the text training data.

Corpus	Frames	Aug	Snt
MuST-C	211M	599M	0.35M
Europarl	30M	80M	0.035M
CoVoST2	711M	202M	1.42M
LibriSpeech	131M	372M	0.1M
TED-LIUM3	163M	463M	0.26M
VoxPopuli	191M	543M	0.18M

Table 3: Statistics of raw and augmented speech corpora. Frames is the audio frames number of the raw data, and Aug is for the audio augmented data. Snt refers to the number of sentences corresponding to the raw audio data.

2.2 Text-to-Text Augmentation

For text-to-text machine translation, augmented data from monolingual corpora in source and target language are generated by knowledge distillation (Kim and Rush, 2016; Freitag et al., 2017) and back translation (Edunov et al., 2018) respectively. Moreover, we use automatic speech recognition (ASR) output utterances to improve MT's robustness.

Back-Translation Back-translation (Sennrich et al., 2015; Lample et al., 2017) is an effective way to improve the translation quality by leveraging a large amount of monolingual data and it has been widely used in WMT campaigns. In our setting, we add a "<BT>" tag to the source side of back-translated data to prevent overfitting on the synthetic data, which is also known as tagged back-translation (Caswell et al., 2019; Marie et al., 2020; Tong et al., 2021).

Knowledge Distillation Sequence-level knowledge distillation (Wang et al., 2021; Sun et al., 2020) is another useful technique to improve translation performance. We enlarge the training data by translating English sentences to Chinese using a good teacher model. Specifically, we trained an EN→ZH offline model based on the deep Transformer as a teacher model. And the beam-search strategy of beam-size 5 is used when translating the English source text into the Chinese target text.

ASR Output Adaptation Traditionally, the

output of ASR systems is lowercased with no punctuation marks, while the MT systems receive natural texts. In our system, we attempt to make our MT systems robust to these irregular texts. A simple method is to apply the same rules on the source side of the MT training set. However, empirical study shows this method causes translation performance degradation. Inspired by the tagged back-translation method (Caswell et al., 2019), we enhance the regular MT models with transcripts from both ASR systems and ASR datasets. An extra tag "<ASR>" indicates the irregular input. Note that the basic idea to bridge the gap between the ASR output and the MT input involves additional sub-systems, like case and punctuation restoration. In our cascaded system, we prefer to use fewer sub-systems, and we will conduct detailed comparison in our future work.

2.3 Speech-to-Text Augmentation

All datasets except MuST-C only contain speech and transcription data. For these datasets, an offline translation model (trained with constrained data) is used to generate Chinese pseudo sentences, which serves as augmented data for training E2E model. In addition, we augment each audio dataset by about 300% using the speed, volume and echo perturbation method as well, and for the CoVoST2 corpus, we augment by 30%. The details are shown in Table 3. Specifically, we first make two copies of all original audio except for CoVoST2. And then, the original audio of all datasets is mixed with all the augmented audio. Finally, we get these training data that are about 1:1 of the original and the augmented audio. Therefore, these training data naturally include the Chinese pseudo data mentioned above. Both ASR and E2E are trained on this training data.

3 Models

3.1 Dynamic-CAAT

Our simultaneous translation systems are based on Cross Attention Augmented Transducer (CAAT) (Liu et al., 2021b), which jointly optimizes the policy and translation model by considering all possible READ-WRITE simultaneous translation action paths. CAAT uses a novel latency loss whose expectation can be optimized by a forward-backward algorithm. Training with this latency loss ensures the controllable latency of CAAT simultaneous translation model. For speech-to-text task, CAAT process the streaming encoder for speech data by block processing with the right context and infinite left context. For text-2-text task, CAAT use conventional unidirectional transformer encoder for text data, which masking the self-attention to only consider previous time-steps.

We improve the training of the CAAT by multiple values of right context window size. Training along multiple right context window size achieves good online performance without setting a prior right context window size in model training. Compared to unidirectional encoder, models trained in this manner can use more source information. The encoder updates the encoder states when new source tokens are available, so that both the encoding of past source tokens and new source tokens are updated. We also show that it is possible to train a single model that is effective across a large range of latency levels.

3.2 Pre-trained LM

For ASR, great advances can be made through pre-training a language model (LM), such as BERT (Devlin et al., 2018), by using sufficient target-domain text (Gao et al., 2021). Inspired by these work, we re-train two language models based on BERT: an English LM and a Chinese LM, respectively for ASR and E2E. Unlike traditional BERT, these two LMs are unidirectional and can be regarded as a special predictor architecture of CAAT.

3.3 Text-to-Text Simultaneous Translation

Our text-to-text Simultaneous Systems are based on Dynamic-CAAT. We use the Dynamic-CAAT implemented based on Transformer, by dividing Transformer's decoder into predictor and joiner module. The predictor and joiner share the same number of transformer blocks as the conventional transformer decoder, while there are no cross-attention blocks in the predictor module and no self-attention blocks in the joiner module.

3.4 Speech-to-Text Simultaneous Translation

3.4.1 Cascaded Systems

The cascaded system includes two modules, simultaneous ASR and simultaneous text-to-text MT. Simultaneous MT system is built with Dynamic-CAAT proposed in Sec. 3.1. However, ASR system directly uses the original CAAT framework for training.

We adjust the range of AL through three hyper-parameters: K, B and P. Where K means the number of ASR output tokens is at least K more than the number of MT output tokens. B is the beam width of MT. P means that the probability of the token generated by the MT model must be greater than P.

The pre-trained LM for ASR is retrained by only using English text corpora descirbed in Sec. 2.1.

3.4.2 E2E Systems

E2E model is built on the original CAAT model. First, we train the E2E model with mixed real and pseudo paired speech-translation data and the scale of the pseudo data is about 1:1 to the real data. Second, pre-training ASR encoder and pre-training LM predictor are used to improve performance under restricted resources. Finally, we train E2E model using multitask learning (Wang et al., 2020a; Ma et al., 2020b; Tang et al., 2021), but didn't achieve the expected effect in this task.

Compared with the tens of millions of data in the MT model, the training data for E2E system is insufficient. So we just train E2E model with low regime, and the E2E model is only used to verify the effectiveness of the training methods.

4 Experiments

In our experiments, pre-norm Transformer-base(Xiong et al., 2020) is used as the offline baseline model to compare with the text-to-text models. The baseline model has 12 encoder layers and 6 decoder layers and it is trained only using bilingual data. We compare the baseline model with three text-to-text models: wait-k(Elbayad et al., 2020), efficient wait-k and Dynamic-CAAT. For speech-to-text task, we compare the results of ASR cascaded Dynamic-CAAT and efficient wait-k respectively. The details of models are summarized in Table 4.

Systems are evaluated with respect to quality and latency. Quality is evaluated with the standard BLEU metric (Papineni et al., 2002). Latency is evaluated with metric average lagging (AL), which is extended to the task of simultaneous speech translation from simultaneous machine translation (Ma et al., 2020d). We conduct all our experiments using Simuleval toolkit (Ma et al., 2020a) and report results for the submitted speech translation tasks. Latest 6 checkpoints of a single training process are averaged in our experiments. We also adopted FP16 mix-precision training to accelerate the training process with almost no loss in BLEU. All models are trained on 8 RTX A10 GPUs. All translation systems are followed by a post-processing module for Chinese punctuation.

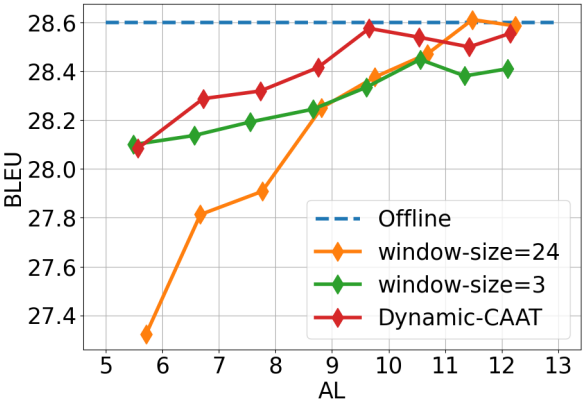

Figure 1: Effectiveness of Dynamic-CAAT

4.1 Effectiveness of Dynamic-CAAT

To demonstrate the effectiveness of Dynamic-CAAT, we compare it with CAAT with different right context window size. Offline results are used for reference, and the offline model has a latency of $AL = |x|$. Models are trained with a batch size of 32,000 token. Figure 1 presents the performance of models trained for a single right context window size w, with $w_{train} \in \{3, 24\}$. Each model is evaluated across different right context window size w, $w_{eval} \in \{4, 5, ..., 11\}$. From Figure 1 we observe that performance of model with $w = 24$ is worse than that of model with right window size $w = 3$, especially $w_{eval} \in \{4, 5, 6\}$). Meanwhile, we found that training on a small right context window size $w = 3$ can generalize well to other w. We note that jointly training on Dynamic right context window size w outperforms training on a single path.

4.2 Effectiveness of Pre-trained LM

We compare the results of the ASR and E2E systems with their respective LM methods. The implementation of our models are based on the CAAT code [1]. For both ASR and E2E tasks, we use specAugment (Park et al., 2019) with $F = 15, m_F = 2, T = 70, p = 0.2, m_T = 2$, and use Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.98$. We set max tokens as 20000

[1] https://github.com/danliu2/caat

Model	Encoder Layers	Decoder Layers/ Predictor Layers	Joiner Layers	Hidden Size	FFN
Offline	12	6	-	512	2048
wait-k	6	6	-	512	1024
efficient wait-k	6	6	-	1024	4096
Dynamic-CAAT	12	6	6	512	2048
ASR	12	6	6	512	2048
E2E	12	6	6	512	2048

Table 4: The details of several model architectures we used.

Models	tst-COMMON (WER / AL)	dev (WER / AL)
ASR-base	13.81 / 927	14.98 / 883
+LM	11.32 / 901	13.32 / 869
Models	**tst-COMMON (BLEU / AL)**	**dev (BLEU / AL)**
E2E-base	19.56 / 1304	17.62 / 1381
+LM	22.49 / 1272	19.71 / 1347

Table 5: Effectiveness of pre-trained LM.

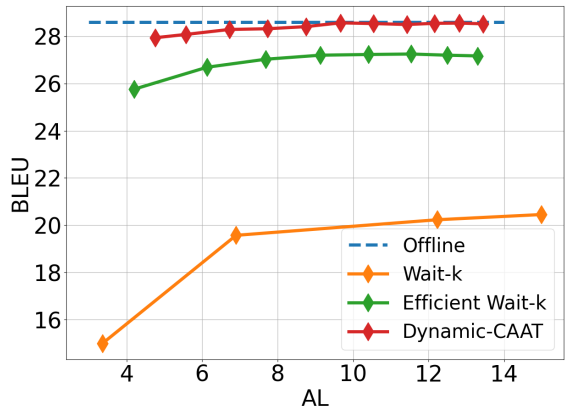

Figure 2: Latency-quality trade-offs of text-to-text simultaneous translation.

and update frequency as 8 during training. And during inference, the beam width is set to 5. Table 5 shows ASR and E2E experiment results. We observe that the ASR and E2E both outperform the baseline systems trained without pre-trained LM.

4.3 Text-to-Text Simultaneous Translation

In text-to-text simultaneous translation task, experiments are conducted on tst-COMMON test set. The latency is measured with the subword-level latency metric. We compare Dynamic-CAAT models with wait-k and efficient wait-k[2]. The results of

[2]https://github.com/elbayadm/attn2d

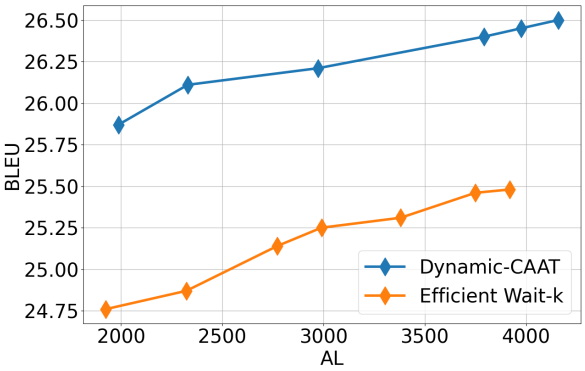

Figure 3: Latency-quality trade-offs of speech-to-text simultaneous translation.

text-to-text EN→ZH are shown in Figure 2. We can see that performance of Dynamic-CAAT is always better than that of wait-k and efficient wait-k, especially in low latency regime, and performance of Dynamic-CAAT is nearly equivalent to offline result.

And during inference, the "<ASR>" tag is added to the front of the ASR output and it can increase 0.2 bleu. For the text-to-text task, we set the beam width to 1.

4.4 Cascaded Speech translation

Under the cascaded setting, we paired two well-trained ASR and Dynamic-CAAT systems. The WER of ASR system's performance is 11.32 with 901 AL, and the cascaded system's results vary with the Dynamic-CAAT hyperparameters K, B, P. The range of K is 3 to 20. P is set to 0.35, and B is set to 1, however, when K is greater than 14, B is set to 6. For comparison, we use another text-to-text machine translation model, efficient wait-k. Performance of cascaded systems is shown in Figure 3. On the test set tst-COMMON from MuST-C v2, the cascaded system of Dynamic-CAAT achieves 25.87, 26.21, 26.45 BLEU with

1987, 2972, 3974 AL respectively. We also find that the BLEU value of Dynamic-CAAT is on average 1.0 higher than that of efficient wait-k in the same AL range.

5 Related Work

5.1 Data Augmentation

In terms of data scale, the amount of training data for speech translation is significantly smaller than that for text-to-text machine translation, and lack of data decreases performance of speech translation. As described in Section 2, based on the text-to-text MT model, sequence-level knowledge distillation and self-training are used to solve the problem of low performance of the speech translation model. This approach has also proven to be the most efficient way to utilize large amounts of ASR training data (Pino et al., 2020; Gaido et al., 2020). In addition, generating speech synthetic data is also effective for low-resource speech recognition tasks (Bansal et al., 2018; Ren et al., 2020).

5.2 Simultaneous Translation

Recent work on simultaneous translation (including S2T and T2T) can be roughly divided into two categories. The first category is represented by the wait-k method, which uses a fixed strategy for the READ/WRITE actions of simultaneous translation, and these models are easy to implement. The second category assumes that adaptive policies are superior to fixed policies, because adaptive policies can flexibly balance the tradeoff between translation quality and latency based on current context information. Research in this category includes supervise learning (Zheng et al., 2019), simultaneous translation decoding with adaptive policy (Zheng et al., 2020), and so on. In addition, researchers have also proposed a monotonic attention mechanism optimized for translation and policy for flexible policy, e.g., Monotonic Infinite Lookback (MILk) attention (Arivazhagan et al., 2019) and Monotonic Multihead Attention (MMA) (Ma et al., 2020c).

6 Conclusion

This paper summarizes the results of the shared tasks in the IWSLT 2022 produced by the AISP-SJTU team. In this paper, Dynamic-CAAT we used outperforms efficient wait-k, and its result is close to offline model in the case of $AL > 9$. From the experiments we also can see that the pre-trained language model plays a most important role in both ASR and E2E translation. Because of the huge difference in the amount of data, the performance of the E2E system is much lower than that of cascaded system. In the future, we hope to explore more effective data augmentation experiments applied to E2E translation. We hope that our practice can facilitate research work and industrial applications.

References

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z. Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. *arXiv preprint arXiv:1906.05218*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12749–12759.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation. *arXiv preprint arXiv:2005.08595*.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Marco Gaido, Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020. *arXiv preprint arXiv:2006.02965*.

Changfeng Gao, Gaofeng Cheng, Runyan Yang, Han Zhu, Pengyuan Zhang, and Yonghong Yan. 2021. Pre-training transformer decoder for end-to-end asr model with unpaired text data. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, et al. 2019. The niutrans machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266.

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021a. The ustc-nelslip systems for simultaneous speech translation task at iwslt 2021. *arXiv preprint arXiv:2107.00279*.

Dan Liu, Mengge Du, Xiaoxi Li, Ya Li, and Enhong Chen. 2021b. Cross attention augmented transducer networks for simultaneous translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 39–55.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. Simuleval: An evaluation toolkit for simultaneous translation. *arXiv preprint arXiv:2007.16193*.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. Simulmt to simulst: Adapting simultaneous text translation to end-to-end simultaneous speech translation. *arXiv preprint arXiv:2011.02048*.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020c. Monotonic multihead attention. In *International Conference on Learning Representations*.

Xutai Ma, Juan Miguel Pino, and Philipp Koehn. 2020d. Simulmt to simulst: Adapting simultaneous text translation to end-to-end simultaneous speech translation. *CoRR*, abs/2011.02048.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, et al. 2020. Wechat neural machine translation systems for wmt20. *arXiv preprint arXiv:2010.00247*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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*.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. *arXiv preprint arXiv:2006.02490*.

Mirco Ravanelli, Titouan Parcollet, and Yoshua Bengio. 2019. The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6465–6469. IEEE.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Simulspeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

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

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

Tzu-Wei Sung, Jun-You Liu, Hung-yi Lee, and Lin-shan Lee. 2019. Towards end-to-end speech-to-text translation with two-pass decoding. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7175–7179. IEEE.

Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel. 2021. A general multi-task learning framework to leverage text data for speech to text tasks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6209–6213. IEEE.

Chengqi Zhao Zhicheng Liu Jian Tong, Tao Wang Mingxuan Wang, Rong Ye Qianqian Dong Jun Cao, and Lei Li. 2021. The volctrans neural speech translation system for iwslt 2021. *IWSLT 2021*, page 64.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020a. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. 2020b. Curriculum pre-training for end-to-end speech translation. *arXiv preprint arXiv:2004.10093*.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming Zhu, Mingxuan Wang, and Lei Li. 2020. The volctrans machine translation system for wmt20. *arXiv preprint arXiv:2010.14806*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. *CoRR*, abs/2002.04745.

Chen Xu, Xiaoqian Liu, Xiaowen Liu, Laohu Wang, Canan Huang, Tong Xiao, and Jingbo Zhu. 2021. The niutrans end-to-end speech translation system for iwslt 2021 offline task. *arXiv preprint arXiv:2107.02444*.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. Wechat neural machine translation systems for wmt21. *arXiv preprint arXiv:2108.02401*.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020. Simultaneous translation policies: From fixed to adaptive. *arXiv preprint arXiv:2004.13169*.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simultaneous translation with flexible policy via restricted imitation learning. *arXiv preprint arXiv:1906.01135*.