Findings of the Third Workshop on Automatic Simultaneous Translation

Ruiqing Zhang[1] **Chuanqiang Zhang**[1] **Zhongjun He**[1] **Hua Wu**[1] **Haifeng Wang**[1]
Liang Huang[2] **Qun Liu**[3] **Julia Ive**[4] **Wolfgang Macherey**[5]

[1] Baidu Inc. [2] Oregon State University [3] Huawei Noah's Ark Lab
[4] Queen Mary University of London [5] Google

Abstract

This paper reports the results of the shared task we hosted on the Third Workshop of Automatic Simultaneous Translation (AutoSimTrans). The shared task aims to promote the development of text-to-text and speech-to-text simultaneous translation, and includes Chinese-English and English-Spanish tracks. The number of systems submitted this year has increased fourfold compared with last year. Additionally, the top 1 ranked system in the speech-to-text track is the first end-to-end submission we have received in the past three years, which has shown great potential. This paper reports the results and descriptions of the 14 participating teams, compares different evaluation metrics, and revisits the ranking method.

1 Introduction

Simultaneous translation (ST), which aims to perform translation from source language speech into the target language with high quality and low latency, is widely used in many scenarios, such as international conferences, live broadcasts, etc.

Generally, the research of ST falls into two categories: the cascade method, and the end-to-end method. A typical cascade ST system consists of an automatic speech recognition (ASR) system that transcribes the source speech into streaming text (Moritz et al., 2020; Wang et al., 2020a; Li et al., 2020a), a machine translation (MT) system that translates the text into the target language, and a policy module lies in between them to decide when to start translation (Oda et al., 2014; Dalvi et al., 2018; Ma et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020; Wilken et al., 2020). Another branch of work proposed end-to-end ST methods that attempt to translate from source speech to target text directly without transcribing the source speech (Bansal et al., 2018; Di Gangi et al., 2019; Jia et al., 2019).

We host a shared task at the Third AutoSimTrans Workshop to promote the exploration of advanced ST approaches. The shared task is built on the past two editions (Wu et al., 2020; Zhang et al., 2021c). We set up three tracks this year:

- **Chinese-English Text-to-text ST track**, where participants are asked to generate real-time English translation based on the input Chinese text. The input is derived from human-annotated transcriptions of TED-like lectures, which contain speech disfluencies but no ASR errors. We simulate streaming speech recognition results by a series of prefixes, where each n-word transcription is represented by n sentence prefixes whose lengths increase from 1 to n.
- **Chinese-English Speech-to-text track** considers real ST scenarios that need real-time translation directly from speech. The participants can adopt either cascade or end-to-end systems. The test sets for the first two tracks are from the same set of audio so that the test results may capture the differences brought by different input modalities.
- **English-Spanish Text-to-text track** is newly added this year. We use the UN Parallel corpus[1] for train and test, which is composed of official records of the United Nations and other parliamentary documents, with no disfluencies and no ASR errors.

The objective of ST systems is to achieve high translation quality with low latency. During the evaluation period, each participant can submit once a day. To examine their quality-latency trade-off ability, the submission of each track is required to contain multiple folders with different policies and varying latency. Our platform supports automatic evaluation and plots the result of each folder to one point on a latency-quality diagram.

[1] https://conferences.unite.un.org/UNCORPUS/en/

Proceedings of the Third Workshop on Automatic Simultaneous Translation, pages 1 - 11

Team	Organization
BIT-Xiaomi	Beijing Institute of Technology & Xiaomi Inc., Beijing, China
Huawei	Huawei Noah's Ark Lab, Guangdong, China
HAU	Huazhong Agricultural University, Hubei, China
USST-ECUST	Univ. of Shanghai for Science and Technology & East China Univ. of Science
HZLHZ	Anonymous
ZXN	Zhejiang Univ. & Xiamen Univ. & North China Institute of Aerospace Engineering
TMU	Tianjin Medical University, Tianjin, China
CITC	Changchun Information Technology College, Jilin, China
NCIAE	North China Institute of Aerospace Engineering, Hebei, China
XJTU	Xi'an Jiaotong University, Shanxi, China
HIT	Harbin Institute of Technology, Heilongjiang, China
ZJU	Zhejiang University, Zhejiang, China
Nuctech	Nuctech Company, Beijing, China
A23	Anonymous

Table 1: List of participants.

We've received 24 submissions from 14 teams this year, 4 times as many as last year. The 14 participants are listed in Table 1. We analyze the submissions and get the following findings:

- The translation quality of the systems, both pipeline and end-to-end in the speech-to-text track lags behind the text-to-text track by more than 9.0 BLEU. This suggests the necessity of exploring robust speech translation systems for pragmatic ST.

- We receive an end-to-end ST submission for the first time in three years, which outperforms all pipeline-based systems submitted this year, representing the potential of end-to-end ST.

- Experiments comparing multiple quality estimation metrics suggest that BLEURT may be more suitable for ST than BLEU given that it correlates best with human ratings.

We will introduce the details of the three tracks (Section 2), report and analyze the submissions (Section 3), and finally compare and analyze evaluation and ranking metrics (Section 4).

2 Shared Task

We first introduce the corpora used in the shared task, then describe the system evaluation method, as well as the differences compared with the past editions.

2.1 Dataset

The corpora provided for training and evaluation are listed in Table 2. For the first two tracks for Zh→En ST, we provide a large-scale text translation corpus, CWMT19[2], along with a speech translation dataset, BSTC (Zhang et al., 2021b). CWMT19 contains 9 million of Zh→En sentence pairs collected from web, bilingual books, movies, law documents, etc. BSTC contains 70.41 hours of Mandarin speeches from three TED-like content producers, corresponding to about 40K source sentences. Compared with last year, we expand the testset of BSTC from 6 talks (1.46 hours) to 20 talks (4.26 hours). For En→Es ST, we use a text translation corpus, the United Nations Parallel Corpus (UN)[3] to simulate the ST scenario. All data can be obtained at the site of our shared task[4] after registration.

The two text-to-text tracks restrict participants to use the provided corpora only, while the speech-to-text track allows the use of additional ASR datasets.

2.2 System Evaluation

The ST systems are evaluated with respect to translation quality and latency. For translation quality, BLEU (Papineni et al., 2002) is the most commonly used metric. Although some net-based approaches such as BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) have been

[2] http://mteval.cipsc.org.cn:81/agreement/description

[3] https://conferences.unite.un.org/UNCORPUS/en

[4] https://aistudio.baidu.com/aistudio/competition/detail/148/

	Corpus	Subset	Talks	Utterances	Transcription (words)	Translation (words)	Audio (hours)
Zh-En	BSTC (ST)	Train	215	37,901	1,004,128	620,263	64.57
		Dev	16	956	24,711	15,794	1.58
		Test	20	2,305	72,695	42,836	4.26
	CWMT19 (MT)	Train	/	9,023,456	264,652,945	182,840,035	/
En-Es	UN (MT)	Train	/	21,911,121	517,327,737	608,514,316	/
		Dev	/	500	12,400	14,701	/
		Test	/	500	13,421	15,935	/

Table 2: The summary of our provided corpora. We calculate the number of talks (Talks), number of sentence pairs (Utterances), number of words[5] in transcription and translation, and the duration of the speeches in corresponding corpora.

proven to be superior to BLEU in text translation, little work has conducted experiments or used them to evaluate ST systems. For the evaluation of latency, recent work have proposed some metrics like Average Proportion (AP) (Cho and Esipova, 2016), Average Lagging (AL) (Ma et al., 2019), Consecutive Wait (CW) (Gu et al., 2017) and Differentiable Average Lagging (DAL) (Arivazhagan et al., 2019).

In our shared task this year, we adopt AL-BLEU and CW-BLEU to evaluate systems in the text-to-text tracks and the speech-to-text track, respectively. AL takes the number of words that the target lags behind the source speaker to estimate the degree of delay. It simulates an ideal policy that generates translation at the same speed as the speaker's utterance and measures the average number of words that lags behind this ideal policy. CW measures the average duration between every two WRITE operations by calculating the average number of source words being waited for.

We will conduct experiments and discuss alternative metrics for evaluating translation quality and latency in Section 4.

2.3 Submission and Ranking

Submission: Each team can participate in multiple tracks. Participants in each track are ranked independently. Different from previous editions, the input of the testsets this year is no longer invisible. Participants only need to submit the simultaneous translation results of the testset to our platform, rather than Docker projects. Before the final submission, participants can submit once a day to view their results and those of other teams on the leaderboard. Each submission needs to contain N ($N \geq 1$) folders containing the ST results with different policies or models. The submissions will be evaluated automatically and plotted to N points on the latency-quality graph. N is determined by the teams themselves.

Ranking: Intuitively, a system is considered better if it generates higher quality results under the same delay or achieves a lower delay when generating results of the same quality. In the shared task, we rank submitted systems based on the Iterative Monotonic Optimal Sequence (I-MOS) algorithm (Zhang et al., 2021c). It iteratively searches for a monotonic optimal sequence (MOS), which contains the points with the best translation quality at corresponding delays. Teams that have points selected on the MOS in the k^{th} iteration are classified to the k^{th} level, then removed from the candidate teams in the $k+1^{th}$ iteration. All teams of the k^{th} level rank higher than that of the $k+1^{th}$ level. Teams belonging to the same level are ranked according to the proportion of points on the MOS.

2.4 Differences With Past Editions

In addition to setting up a new En-Es text-to-text ST track, this year's shared task has the following two differences compared with the past editions:

- Participants submit ST results instead of docker projects, which is much easier for participants. For this, we released the audios and corresponding transcription for the first two tracks of Zh-En ST and extended the testset from 6 talks to 20.

- This year's shared task allows each team to submit once per day, rather than only once in the entire challenge period. We developed an automated evaluation platform, enabling participants to access their evaluation results in real-time.

[5]Record the number of characters in the Transcriptions for Chinese.

Rank	Team	Score
1	BIT-Xiaomi	7.00
2	Huawei	6.00
2	USST-ECUST	6.00
4	HZLHZ	4.50
4	HAU	4.50
6	TMU	4.00
7	CITC	3.33
8	NCIAE	3.33
9	ZXN	2.67
10	XJTU	2.00
11	HIT2	1.67
12	ZJU	1.50
13	Nuctech	1.00

Table 3: The ranking of the Zh→En text-to-text ST track. The scores are calculated according to the I-MOS algorithm.

Rank	Team	Score
1	Huawei	2.00
2	BIT-Xiaomi	1.50
3	ZXN	1.00
3	HAU	1.00

Table 4: The ranking of the Zh→En speech-to-text ST track.

3 System Results

3.1 Chinese-English Simultaneous Translation

The first two tracks are for Chinese-English ST from Chinese text and speech, respectively. We've received submissions from 13 teams: 13 entered the text-to-text track and 4 of them also participated in the speech-to-text track. Their latency-quality trade-off results are plotted in Figure 1.

3.1.1 The Text-to-text track

The ranking of the 13 participants in the Zh→En text-to-text track is shown in Table 3. We list the approaches used by some of the participants as follows:

- **BIT-Xiaomi** (Liu et al., 2022) changed the granularity in wait-k policy (Ma et al., 2019) from Chinese characters to words. They proposed to train a streaming word segmentation model to detect Chinese word boundaries in real-time, and performed prefix-to-prefix training of wait-k according to the number of words. The MT model is a Transformer-big (Vaswani et al., 2017) model trained with data selection, data augmentation (Sennrich et al., 2015), R-drop (Wu et al., 2021), and noise adding strategies to improve the model's robustness.

- **USST-ECUST** (Zhu and Yu, 2022) adopted the Transformer with 12 encoders and 6 decoders as the MT model, which is pre-trained on a large-scale Zh-En corpus containing 9 million sentence pairs from CWMT19 and 5.7 million pairs of pseudo data generated through self-training (He et al., 2019) and back-translation (Sennrich et al., 2015; Edunov et al., 2018). The model is then fine-tuned with prefix-to-prefix training (Ma et al., 2019) on a mixture of BSTC corpus and a subset of CWMT19 that is most similar to BSTC for better domain adaptation.

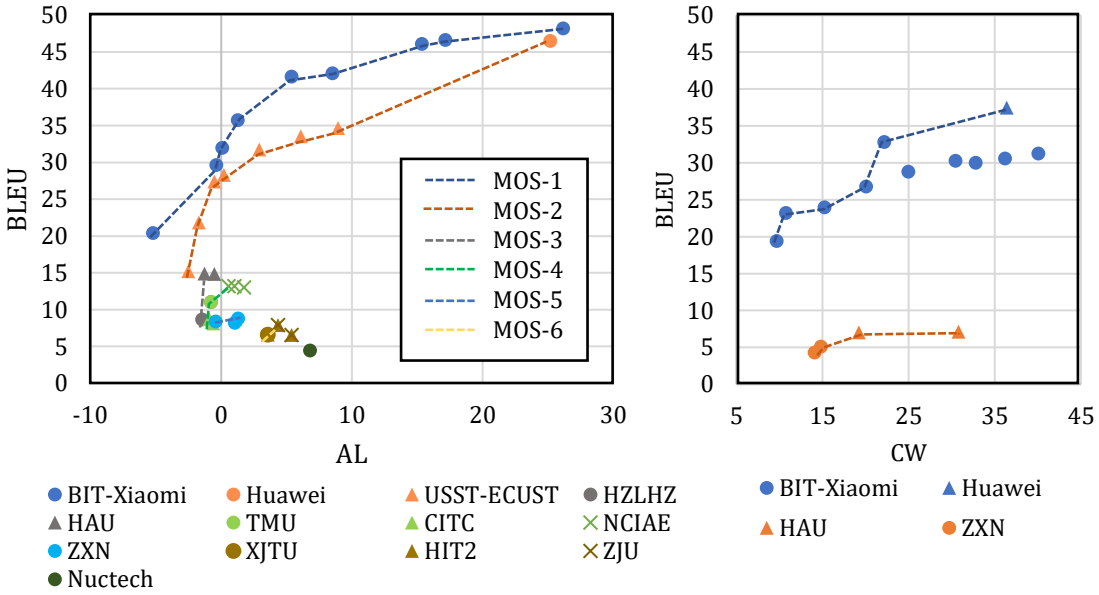
- **HAU** (Zhang, 2022) trained a prefix-to-prefix model using the wait-k policy with $k = 1$ and 3 in the text-to-text simultaneous translation.

3.1.2 The Speech-to-text track

The ranking of the 4 participants in the Zh→En speech-to-text track is listed in Table 4.

- **Huawei** (Zeng et al., 2022) built an end-to-end simultaneous translation model based on RealTranS (Zeng et al., 2021). It includes a CTC-guided acoustic encoder, a semantic encoder, and a translation decoder. The acoustic encoder is initialized from a pre-trained ASR model, and the semantic encoder and the translation decoder are initialized from a pre-trained NMT model. In the fine-tuning stage, they first generated pseudo ST training data by translating the transcripts of 20,000 hours of in-house ASR corpora into the target text, then train the model with the multi-path wait-k (Elbayad et al., 2019) policy on the pseudo data together with BSTC.

- **BIT-Xiaomi** (Liu et al., 2022) took a pipeline system. The audio inputs are firstly segmented by Silero-VAD (Team, 2021), then sent to a Transformer-based ASR model trained on AISHELL-1 (Bu et al., 2017) and BSTC (Zhang et al., 2021b). The recognized text is then sent to the policy model and the MT model to decide when to translate and produce a translation. The MT model and the



(a). Zh-En Text-to-text ST track

(b). Zh-En Speech-to-text ST track

Figure 1: The evaluation results of the first two tracks. The order in the legend (line by line) denotes the ranking result, which is calculated by the I-MOS algorithm. It iteratively builds the monotonic optimal sequence (MOS) of level k (MOS-k) and classifies teams that have points on it to the k^{th} level. We use points of the same color but different shapes to represent the results of teams belonging to the same level, and the teams are ranked according to the proportion of points on the corresponding MOS.

policy module are the same as they used in the text-to-text track.

- **ZXN** (Li et al., 2022) developed a pipeline system with an audio segmentation model, an ASR system, and a wait-k based MT model. The audio segmentation model performs endpoint detection (EPD) based on short-term energy and zero-crossing rate (Rabiner and Sambur, 1975). The ASR system includes a convolutional model with a CTC decoder (Graves et al., 2006) to generate pinyin sequences, followed by a language model based on the maximum entropy markov model (MEMM) to produce Chinese characters. The MT model adopts Transformer-base (Vaswani et al., 2017) and is trained with the prefix-to-prefix mode. The ASR model is pre-trained on AISHELL-1 and Thchs-30 (Wang and Zhang, 2015), and the MT model is pre-trained on CWMT19, then both are fine-tuned on the BSTC.

- **HAU** (Zhang, 2022) also took a pipeline system. They adopted DeepSpeech2 (Amodei et al., 2016) as the ASR model, which is trained on AISHELL-1 only without further fine-tuning on BSTC. The ST policy and the MT model they used are the same as they used in the text-to-text track.

	Text-to-text	Speech-to-text
BIT-Xiaomi	48.17	31.26
Huawei	46.49	37.46

Table 5: The highest BLEU scores achieved by BIT-Xiaomi and Huawei for the same testset with different input modalities. The Speech-to-text track inputs audios while the Text-to-text track inputs golden transcription.

Table 5 lists the highest translation quality achieved by BIT-Xiaomi and Huawei, the two best performing teams on the two tracks. Compared to their performance on the text-to-text track, their speech-to-text systems both have a BLEU degradation of over 9 points. This quality gap is brought about by different input modalities. The speech-to-text systems receive audio as input, so they need an ASR model to transcribe the audio, or an end-to-end speech translation model to generate translation directly from speech. The pipeline systems have the problem of error propagation, and the performance of the end-to-end systems is limited by data scarcity.

This also gives us some hints that the processing of speech may be the most significant factor

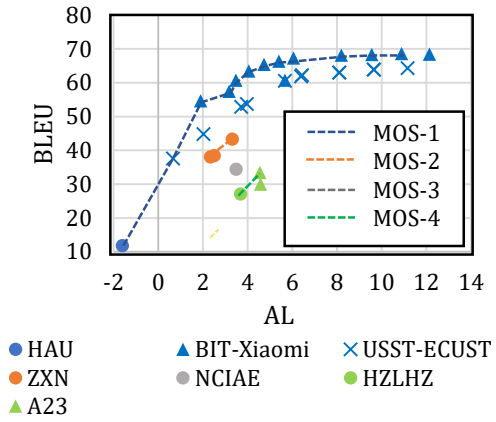


Figure 2: The evaluation results of the En-Es text-to-text ST track.

Rank	Team	Score
1	HAU	4.00
2	BIT-Xiaomi	3.83
3	USST-ECUST	3.08
4	ZXN	3.00
5	NCIAE	2.00
6	HZLHZ	1.00
7	A23	0.50

Table 6: The ranking of the En→Es text-to-text ST track.

affecting the effect of simultaneous translation in real scenes. Some work has attempted to improve the pipeline systems by introducing an ASR error correction model (Leng et al., 2021; Zhang et al., 2021a), others proposed pre-training approaches to alleviate the data scarcity problem of speech translation corpora in end-to-end systems (Wang et al., 2020b; Pino et al., 2020; Zheng et al., 2021; Li et al., 2020b; Zhang et al., 2022). We hope to see more participants in future workshops investigating how to close the performance gap between the two tracks.

3.2 English-Spanish Simultaneous Translation

The En→Es track received submissions from 7 teams. The latency-quality trade-off results of the En-Es track are plotted in Figure 2 and the ranking is listed in Table 6. According to the system descriptions submitted, almost all teams used the same training policies in this track as in the Zh→En text-to-text track.

4 Discussion

We first carry out experiments to compare different translation quality evaluation metrics (Section 4.1), then discuss a controversial ranking dilemma of I-MOS algorithm in the ranking algorithm (Section 4.2).

4.1 BLEU, BERTScore, and BLEURT

	Metrics	$r(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$
SYS1	SentBLEU	0.546	0.484	0.390
	BERTScore	0.553	0.484	0.388
	BLEURT	**0.708**	**0.655**	**0.537**
SYS2	SentBLEU	0.584	0.516	0.415
	BERTScore	0.587	0.540	0.433
	BLEURT	**0.729**	**0.693**	**0.568**
SYS3	SentBLEU	0.525	0.468	0.374
	BERTScore	0.529	0.498	0.396
	BLEURT	**0.670**	**0.654**	**0.532**
SYS4	SentBLEU	0.467	0.408	0.322
	BERTScore	0.135[6]	0.467	0.368
	BLEURT	**0.637**	**0.629**	**0.507**
SYS5	SentBLEU	0.451	0.422	0.332
	BERTScore	0.518	0.522	0.414
	BLEURT	**0.656**	**0.672**	**0.539**
SYS6	SentBLEU	0.370	0.350	0.274
	BERTScore	0.475	0.480	0.376
	BLEURT	**0.559**	**0.578**	**0.459**

Table 7: Sentence-level agreement with human ratings on 6 ST systems. Given 6 source documents, each system (SYSi) performs ST, and the translation results are evaluated by sentenceBLEU (sentBLEU), BertScore, and BLEURT with 4 references. We calculate the Pearson correlation (r), the Spearman correlation (ρ), and the Kendall Tau (τ) score between the automatic metrics and human ratings. BLEURT has obvious advantages over the other two metrics in all the 6 systems.

Recently, many quality estimation metrics have been proposed to better imitate human evaluation (Specia et al., 2021), such as *RUSE* (Shimanaka et al., 2018), *YiSi* (Mathur et al., 2019), *BERTScore* (Zhang et al., 2019), *BLEURT* (Sellam et al., 2020), etc. These metrics are proven to be superior to traditional quality evaluation metrics like *BLEU* (Papineni et al., 2002) in text translation. However, to the best of our knowledge, no work has conducted experiments in the ST scenario, and almost all ST work still takes *BLEU* as the criterion for translation quality evaluation.

To keep consistent with previous work, we still

[6]This outlier is caused by a missing translation (one sentence generates an empty translation). Different from BERTScore, SentBLEU and BLEURT are less influenced because the BERTscores are relatively high (always higher than 0.9), for which one zero brought by empty translation would largely degrade its Pearson correlation score.

used the document-level BLEU[7] for evaluation in the shared task this year. Now we conduct experiments to compare it with sentence-level BLEU, BERTScore[8], and BLEURT[9].

4.1.1 Agreement between automatic metrics and human ratings

To evaluate the SOTA quality estimation metrics, we ask human annotators to assess the results of multiple ST systems and calculate the agreement between automatic metrics and human ratings. Each sentence is rated to 1, 2, or 3. 1 denotes the translation is inconsistent with the original text, or incomprehensible; 2 denotes the translation conveys the main idea of the original text but with minor mistakes in grammar or word usage; 3 denotes the translation is fully consistent with the original text. In order to ensure uniform rating standard, all evaluated sentences are scored by one annotator first, and then checked by another annotator. The two annotators are both translators who graduated from Chinese-English translation major.

We randomly select 6 documents (including 975 source sentences in total) from the testset of the first track for evaluation, and then select 6 ST systems with high BLEU scores on this testset (SYS1: 30.23, SYS2: 30.35, SYS3: 29.38, SYS4: 33.45, SYS5: 42.05, SYS6: 41.27) and have they manually rated. Given the simultaneous translation result produced by 6 systems, we calculate the Pearson correlation (r), the Spearman correlation (ρ), and the Kendall Tau (τ) points between human ratings and scores of different automatic metrics. As shown in Table 7, *BLEURT* has a higher correlation with human ratings compared with the other two metrics in all the 6 systems.

4.1.2 Using different metrics for ranking

Next, we explore these metrics from the perspective of ranking. Taking the average score of all the evaluated sentences as the ranking basis, we wonder whether each metric would yield a ranking consistent with human evaluations. We first count the proportion of sentences with a human rating of 2 or 3 as the acceptability for each system. Figure 3 shows that the rank (horizontal axis) of the six systems in terms of acceptability, from

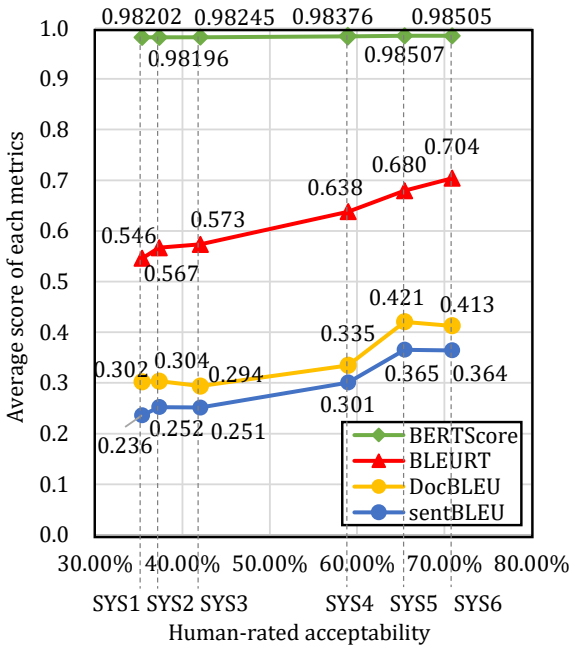


Figure 3: Human-rated acceptability vs. automatic metrics for the translation of 6 systems.

Metrics	$r(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$
DocBLEU	0.917	0.771	0.600
SentBLEU	0.970	0.886	0.733
BERTScore	0.968	0.886	0.733
BLEURT	**0.994**	**1.000**	**1.000**

Table 8: Document-level agreement with human ratings.

low to high is: SYS1 < SYS2 < SYS3 < SYS4 < SYS5 < SYS6. Comparing the human-rated acceptability scores and the quality estimated by automatic metrics, we find that Document-level BLEU (*DocBLEU*) and Sentence-level BLEU (*sentBLEU*) score SYS3 inferior to SYS2, *BERTScore* rates SYS2 inferior to SYS1, and all the three metrics rank SYS6 inferior to SYS5. The ranking results of all the three metrics are different from those given by the human-rated acceptability. On the contrary, *BLEURT*'s ranking for the 6 systems is consistent with the human results, indicating its higher accuracy in imitating human judgment. Note that, *BERTScore* rates all systems around 0.98, with no significant differences. This might be caused by the collapse problem (Chen and He, 2021; Yan et al., 2021), meaning that BERT-derived representations are somehow collapsed, so that almost all sentences are mapped to a similar representation and produce high similarity.

Table 8 further lists the correlation between the automatic metrics and human acceptability for the 6

[7] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

[8] https://github.com/Tiiiger/bert_score based on roberta-large

[9] https://github.com/google-research/bleurt with BLEURT-20

	Metrics	$r(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$
SYS3	BLEURT	0.642	0.604	**0.528**
	+ft	**0.654**	**0.620**	0.502
SYS5	BLEURT	0.590	0.597	0.484
	+ft	**0.703**	**0.704**	**0.569**
SYS6	BLEURT	0.526	0.544	0.439
	+ft	**0.639**	**0.643**	**0.516**

Table 9: The correlation between human ratings and BLEURT scores, before and after fine-tuning.

systems, demonstrating the superiority of *BLEURT* to all the other three metrics.

4.1.3 Fine-tuning *BLEURT* on human annotations

We further attempt to improve the performance of *BLEURT* by fine-tuning on some human ratings. We first construct a quality estimation training set consisting of $975 \times 3 \times 4 = 11700$ triples <hypo, ref, score> built by pairing the ST results (hypo) and human ratings (score) of three systems (SYS1, SYS2, and SYS4) with corresponding 4 references (ref). Then we fine-tune *BLEURT* on this training set and evaluate its performance on the remaining three systems. Here we use *BLEURT-Base*[10] for faster training.

The improvements brought by fine-tuning is shown in Table 9. After fine-tuning, the correlation of almost all systems has been significantly improved, especially for SYS5 and SYS6.

4.2 The ranking dilemma

In the shared task, we take the I-MOS algorithm for ranking. It iteratively builds a monotonic optimal sequence (MOS) and considers the proportion of optimal points as the ranking basis. On the quality-latency figure, the MOS is a sequence of optimal points with increasing translation quality and latency, and a point is considered optimal if there is no other point or line above it at an identical latency. Although I-MOS is adaptive to uncertain submission results, it has one drawback, that is, the MOS curve is bound to select the leftmost point regardless of its translation quality, because the leftmost point is definitely an optimal point. Therefore, I-MOS somehow encourages participants to submit only one point with extremely low latency, making the team ranked first place by the I-MOS algorithm, the leftmost point of Figure 2 is such a case.

[10] https://storage.googleapis.com/bleurt-oss/bleurt-base-128.zip

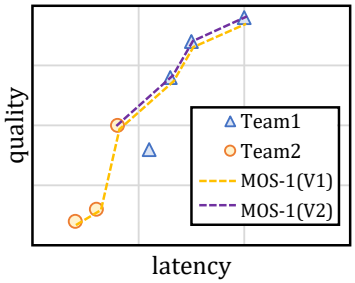


Figure 4: An example to illustrate the ranking dilemma of the I-MOS ranking algorithm. The vanilla I-MOS algorithm calculates MOS-1 as the yellow dotted curve (V1). According to V1, Team2 would rank higher than Team1, although its left two points are unconvincing because of their extremely low quality. After applying our proposed remedy, the left two points of Team2 are removed and Team1 ranks higher based on the modified MOS-1(V2).

To eliminate the defect of I-MOS, we propose to add two strategies to future shared tasks:

1. We require each team to submit at least two points with different delays to make a latency-quality trade-off.

2. Before running the I-MOS algorithm, we first scan to remove the leftmost points whose quality is worse than others' submissions. If all submission points of a team are removed, the team will be ranked last.

   See Figure 4 for example. The vanilla I-MOS algorithm would generate the dashed curve as MOS-1 (V1), causing Team2 to rank higher (Team1 scores $3/4$, Team2 scores $3/3$), although its left two points are unconvincing due to their extremely low quality. But after applying this strategy, we will remove the two points of Team2 because no other team has points with inferior quality compared to them. Then Team2 will be scored to $1/3$. We don't have to worry whether this strategy will lead to unfairness if Team2 is designed for ST at low latency. If Team2 doesn't deliberately take advantage of the defect of I-MOS, they should submit more results at higher latency, at least submit their full-sentence translation result.

5 Conclusion and Future work

This paper presents the results of the simultaneous translation shared task we hosted at the 3^{rd} Workshop on Automatic Simultaneous Translation work-

shop. The shared task includes three tracks, two text-to-text tracks in different languages, and one speech-to-text track. We analyze the submissions from 14 participating teams and have the following inspirations for future ST work:

1. **Robust ST model**: The results of the first two tracks reveal there exists a great gap between using speech input and its corresponding golden transcriptions. Therefore, it is important to explore robust speech translation systems in real ST scenes.

2. **End-to-end ST**: In the speech-to-text track, we received an end-to-end ST submission system for the first time in three years. It integrates a read-write policy into an end-to-end speech translation model and outperforms all the cascaded systems, representing the potential of end-to-end simultaneous translation models.

3. **Quality Evaluation**: Although recently proposed neural network-based metrics are proven superior to BLEU for standard text translation, ST work always takes BLEU for quality estimation. We compare multiple metrics under the ST scenario and verify that BLEURT is more suitable than BLEU for ST in terms of correlation with human ratings.

4. **System Evaluation**: We propose the I-MOS algorithm as well as its revised version for system ranking. Considering both quality and latency is crucial for a practical ST system. However, the quality-latency metric for ST systems is rarely studied. We suggest further study on this topic.

In future shared tasks, we will make the following changes:

1. **Submission**: Add a requirement that each submission should contain at least two points with different delays to make a latency-quality trade-off.

2. **Criterion**: Use BLEURT to replace BLEU for its better correlation with human ratings.

3. **Ranking**: Removing the leftmost points whose quality is worse than others' submission before running the I-MOS algorithm.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

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.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499, New Orleans, Louisiana. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Dessi Roberto, and Marco Turchi. 2019. Enhancing transformer for end-to-end speech-to-text translation. In *Machine Translation Summit XVII*, pages 21–31. European Association for Machine Translation.

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. 2019. Efficient wait-k models for simultaneous machine translation. In *Interspeech*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE.

Yichong Leng, Xu Tan, Rui Wang, Linchen Zhu, Jin Xu, Wenjie Liu, Linquan Liu, Tao Qin, Xiang-Yang Li, Edward Lin, et al. 2021. Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition. *arXiv preprint arXiv:2109.14420*.

Bo Li, Shuo-yiin Chang, Tara N Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu. 2020a. Towards fast and accurate streaming end-to-end asr. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6069–6073. IEEE.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020b. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

Zecheng Li, Yue Sun, and Haoze Li. 2022. System description on automatic simultaneous translation workshop. In *The 3rd Workshop on Automatic Simultaneous Translation at NAACL 2022*.

Mengge Liu, Xiang Li, Bao Chen, Yanzhi Tian, Tianwei Lan, Silin Li, Yuhang Guo, Jian Luan, and Bin Wang. 2022. BIT-xiaomi's system for autosimtrans 2022. In *The 3rd Workshop on Automatic Simultaneous Translation at NAACL 2022*.

Mingbo Ma, Liang Huang, Hao Xiong, Kaibo Liu, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, and Haifeng Wang. 2019. STACL: simultaneous translation with integrated anticipation and controllable latency. In *ACL 2019*, volume abs/1810.08398.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808.

Niko Moritz, Takaaki Hori, and Jonathan Le. 2020. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 551–556.

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*, pages 311–318. Association for Computational Linguistics.

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*.

Lawrence R Rabiner and Marvin R Sambur. 1975. An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2):297–315.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

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

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. Association for Computational Linguistics.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Chengyi Wang, Yu Wu, Shujie Liu, Jinyu Li, Liang Lu, Guoli Ye, and Ming Zhou. 2020a. Low latency end-to-end streaming speech recognition with a scout network. *arXiv preprint arXiv:2003.10369*.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

Dong Wang and Xuewei Zhang. 2015. Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882*.

Patrick Wilken, Tamer Alkhouli, Evgeny Matusov, and Pavel Golik. 2020. Neural simultaneous speech translation using alignment-based chunking. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 237–246, Online. Association for Computational Linguistics.

Hua Wu, Collin Cherry, Liang Huang, Zhongjun He, Mark Liberman, James Cross, and Yang Liu, editors. 2020. *Proceedings of the First Workshop on Automatic Simultaneous Translation*. Association for Computational Linguistics, Seattle, Washington.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Xingshan Zeng, Liangyou Li, and Qun Liu. 2021. Realtrans: End-to-end simultaneous speech translation with convolutional weighted-shrinking transformer. *arXiv preprint arXiv:2106.04833*.

Xingshan Zeng, Pengfei Li, Liangyou Li, and Qun Liu. 2022. End-to-end simultaneous speech translation with pretraining and distillation: Huawei noah's system for autosimtrans 2022. In *The 3rd Workshop on Automatic Simultaneous Translation at NAACL 2022*.

Ruiqing Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Learning adaptive segmentation policy for end-to-end simultaneous translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7862–7874, Dublin, Ireland. Association for Computational Linguistics.

Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021a. Correcting Chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2250–2261, Online. Association for Computational Linguistics.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021b. BSTC: A large-scale Chinese-English speech translation dataset. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35, Online. Association for Computational Linguistics.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2021c. Findings of the second workshop on automatic simultaneous translation. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 36–44, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yiqiao Zhang. 2022. System description on third automatic simultaneous translation workshop. In *The 3rd Workshop on Automatic Simultaneous Translation at NAACL 2022*.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*, pages 12736–12746. PMLR.

JiaHui Zhu and Jun Yu. 2022. USST's system for autosimtrans 2022. In *The 3rd Workshop on Automatic Simultaneous Translation at NAACL 2022*.