BSTC: A Large-Scale Chinese-English Speech Translation Dataset

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He
Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, Qinfei Li
Baidu Inc. No. 10, Shangdi 10th Street, Beijing, 100085, China
{zhangruiqing01, zhangchuanqiang, hezhongjun, wu_hua}@baidu.com

Abstract

This paper presents BSTC (Baidu Speech Translation Corpus), a large-scale Chinese-English speech translation dataset. This dataset is constructed based on a collection of licensed videos of talks or lectures, including about 68 hours of Mandarin data, their manual transcripts and translations into English, as well as automated transcripts by an automatic speech recognition (ASR) model. We have further asked three experienced interpreters to simultaneously interpret the testing talks in a mock conference setting. This corpus is expected to promote the research of automatic simultaneous translation as well as the development of practical systems. We have organized simultaneous translation tasks and used this corpus to evaluate automatic simultaneous translation systems.

1 Introduction

In recent years, automatic speech translation (AST) has attracted increasing interest for its commercial potential (*e.g.*, *Simultaneous Interpretation* and *Wireless Speech Translator*). A large amount of research has focused on speech translation (Weiss et al., 2017; Niehues et al., 2018; Chung et al., 2018; Sperber et al., 2019; Kahn et al., 2020; Inaguma et al., 2020) and simultaneous translation (Sridhar et al., 2013; Oda et al., 2014; Cho and Esipova, 2016; Gu et al., 2017; Ma et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020). The former intends to convert speech signals in the source language to the target language, and the latter aims to achieve a real-time translation that delivers the speech to the audience in the target language while minimizing the delay between the speaker and the translation.

To train an AST model, existing corpora can be classified into two categories:

- ***Speech Translation*** corpora consist pairs of audio segments and their corresponding translations.
- ***Simultaneous Translation*** corpora are constructed by transcribing lecturers' speeches and the streaming utterance of human interpreters.

Speech Translation	**Languages**	**Hours**
F-C (2013)	Es→En	38
KIT-Disfluency (2014)	De→En	13
BTEC (2016)	En→Fr	17
MSLT V1.0 (2016)	En↔Fr/De	23
	En→Zh/Jp	6
MSLT V1.1 (2017)	Zh→En	5
	Jp →En	9
Travel (2017)	Am→En	8
Aug-LibriSpeech (2018)	En→Fr	236
MuST-C (2019)	En→8 Euro langs	3617
Europarl-ST (2020)	9 Euro langs	1642
Covost (2020a; 2020b)	En↔21 langs	2880
Simultaneous Translation	**Languages**	**Hours**
CIAIR (2004)	En↔Jp	182
EPPS (2009)	En↔Es	217
Simul-Trans (2014)	En↔Jp	22
BSTC (ours)	Zh→En	68

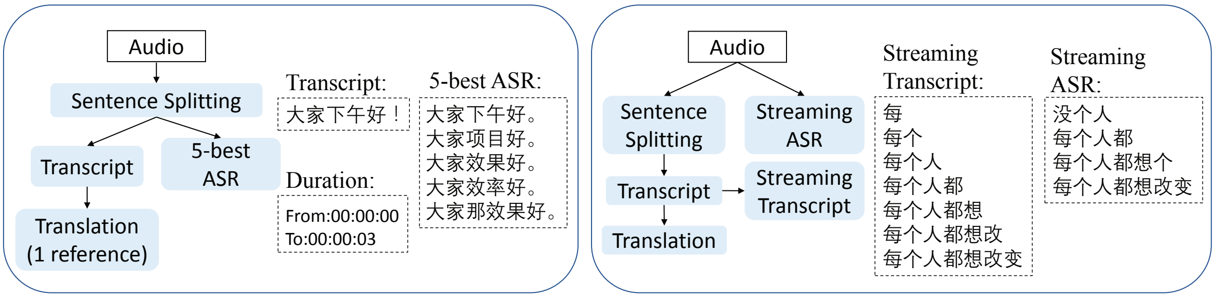
Table 1: Existing speech translation corpora and ours. The duration statistics of all datasets are rounded up to an integer hour. For MuST-C, the "8 Euro langs" is short for "8 European languages". Europarl-ST contains the speech translation between 9 European languages.

The main difference between these two kinds of corpora lies in the way that the translations are generated. The translations in *Speech Translation* corpora are generated based on complete audios or their transcripts, while the translations in *Simultaneous Translation* corpora are transcribed from real-time human interpretation.

Existing research on *Speech Translation* mainly focused on the translation between English and Indo-European languages[1], with little attention paid to that between Chinese (Zh) and English. One of the reasons is the scarcity of public Zh↔En

[1] Indo-European languages are a large language family.

Proceedings of the 2nd Workshop on Automatic Simultaneous Translation, pages 28–35
June 10, 2021. ©2021 Association for Computational Linguistics



(a). Training set

(b). Development set and Testset

Figure 1: The process of constructing the training set and development/test sets (dev/test). The difference between the two processes is that for the training set we first split audio into sentences and then get the ASR and transcript for each sentence, while for the dev/test sets we record the real-time ASR and transcript, the sentence splitting is only used to generate translations of segmented sentences.

speech translation corpora. Among the public corpora, only MSLT (Federmann and Lewis, 2017) and Covost (Wang et al., 2020a,b) contains Zh↔En speech translation, as shown in Table 1. But the total volume of them on Zh→En translation is merely about 30 hours, which is too small to train data-hungry neural models. Some studies explore Zh→En *Simultaneous Translation* (Ma et al., 2019; Zhang et al., 2020). However, they take text translation datasets to simulate real-time translation scenarios because of the lack of simultaneous translation corpus.

To promote the research on Chinese-English speech translation, as well as evaluating the translation quality in real simultaneous interpretation environments, we construct BSTC, a large-scale Zh→En speech translation and simultaneous translation dataset including approximately 68 hours of Mandarin speech data with their automatic recognition results, manual transcripts, and translations. Our contributions are:

- We propose the first large-scale (68 hours) Chinese-English *Speech Translation* corpus. This training set is a four-way parallel dataset of Mandarin audio, transcripts, ASR lattices, and translations.
- The proposed dev and test set constitutes the first high-quality *Simultaneous Translation* dataset of over 3-hour Mandarin speech, together with its streaming transcript, streaming ASR results, and high-quality translation.
- We have organized two simultaneous interpretation tasks[2] to promote research in this field and deployed a strong benchmark on this dataset.
- The proposed dataset can also be taken as 1) a *Chinese Spelling error Correction* (CSC) corpus containing pairs of ASR results and corresponding manual transcripts or 2) a Zh→En *Document Translation* dataset with context-aware translations.

2 Dataset Description

BSTC is created to fill the gap in Zh→En speech translation, in terms of both size and quality. To achieve these objectives, we start by collecting approximate 68 hours of mandarin speeches from three TED-like content producers: BIT[3], *tndao.com*[4], and *zaojiu.com*[5]. The speeches involve a wide range of domains, including IT, economy, culture, biology, arts, etc. We randomly extract several talks from the dataset and divide them into the development and test set.

2.1 Training set

For the training set, we manually tag timestamps to split the audio into sentences, transcribe each sentence and ask professional translators to produce the English translations. The translation is generated based on the understanding of the entire talk and is faithful and coherent as a whole. To facilitate the research on robust speech translation, we also provide the top-5 ASR results for each segmented speech produced by SMLTA[6], a streaming multi-

[2] We organized two shared tasks on the 1st and 2nd Workshop on Automatic Simultaneous Translation.

[3] https://bit.baidu.com

[4] http://www.tndao.com/about-tndao

[5] https://www.zaojiu.com/

[6] http://research.baidu.com/Blog/index-view?id=109

Dataset	Talks	Utterances	Transcription (characters)	Translation (tokens)	Audio (hours)	WER(1-best)
Train	215	37,901	1,028,538	606,584	64.57	27.90%
Dev	16	956	26,059	75,074	1.58	15.21%
Test	6	975	25,832	70,503	1.46	10.32%

Table 2: The summary of our proposed speech translation data.

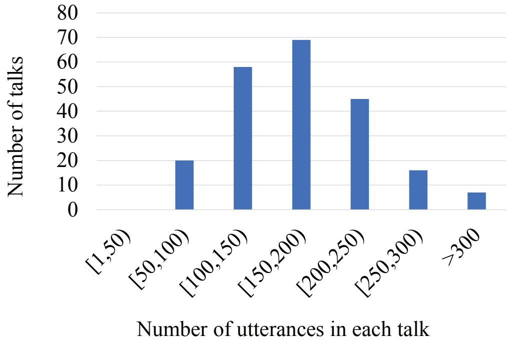


Figure 2: The distribution of talk length (number of sentences) in the training set.

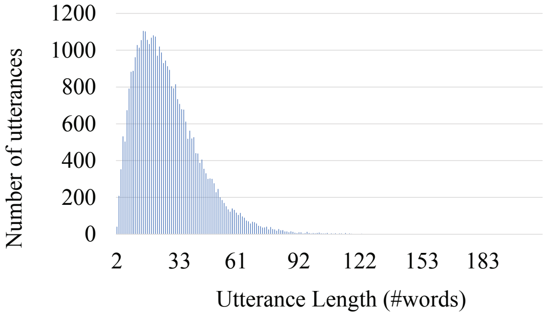


Figure 3: The distribution of utterance length (number of words) in the training set. A word means a Chinese character here.

layer truncated attention ASR model. Figure 1 (a) shows the construction process of the training set, together with an example of a segmented sentence.

2.2 Dev/Test set

For the development (dev) set and test set, we consider the simultaneous translation scenario and provide the streaming transcripts and streaming ASR results, as shown in Figure 1 (b). The streaming transcripts are produced by turning each n-words (a word means a Chinese character here) sentence to n lines word by word with length $1, 2, ..., n$. We use the real-time recognition results of each speech, rather than the recognition of each sentence-segmented audio. This is to simulate the simultaneous interpreting scenario, in which the input is streaming text, rather than segmented sentences.

d_{len}	WER	Coverage
0	5.87%	31.61%
1	7.13%	55.30%
3	8.86%	68.50%
7	10.72%	74.50%
15	15.23%	83.40%
31	23.51%	94.00%
∞	27.90%	100%

Table 3: The WER and coverage of different subsets of the training set with the length difference Δ_{len} between transcript and asr lower than or equal to d_{len}.

2.3 Statistics and Dataset Features

We summarize the statistics of our dataset in Table 2. The distribution of talk length and utterance length in the training set is illustrated in Figure 2 and Figure 3, respectively. The average number of utterances per talk is 176.3 in the training set, 59.8 in the dev set, and 162.5 in the test set. And the average utterance length is 27.14 in the training set, 27.26 in the dev set, and 26.49 in the test set.

We also calculate the word error rate[7] (WER) of the ASR system on the three datasets. As shown in Table 2, the WER of the training set is 27.90%, significantly higher than that of the dev and testset. This is due to the way of audio segmentation before recognition: some audio clips lose some parts in acoustic truncation, resulting in incomplete ASR results. We count the length difference of each <transcription, asr> pair, i.e., $\Delta_{len} = |len(transcription) - len(asr)|$, and recalculate the WER of pairs whose length difference is within a certain range. The WER and coverage of these subsets are listed in Table 3. Note that when the asr and transcript with equal length ($\Delta_{len} \leq 0$), the WER is only 5.87%. For the length difference in a relatively regular range (e.g, $\Delta_{len} \leq 15$), the WER is also relatively low (WER=15.23%).

Besides, there is a difference between our dataset and the existing speech translation corpora. In our dataset, speech irregularities are kept in transcrip-

[7] WER tool: https://github.com/belambert/asr-evaluation

	BLEU	AP	Omissions
A	24.20	83.0%	53%
B	17.14	62.8%	47%
C	25.18	76.5%	53%

Table 4: Comparison of the simultaneous interpretation results of three interpreters (A, B, and C) on the BSTC test set. “AP” is the Acceptability and the “Omissions” indicates the proportion of missing translation in all translation errors.

tion while omitted in translation (eg. filler words like “嗯, 呃, 啊”, unconscious repetitions like “这个这个呢” and some disfluencies), which can be used to evaluate the robustness of the NMT model dealing with spoken language. Some other large-scale speech translation datasets (Kocabiyikoglu et al., 2018; Di Gangi et al., 2019), on the contrary, ignore these speech irregularities in the transcript.

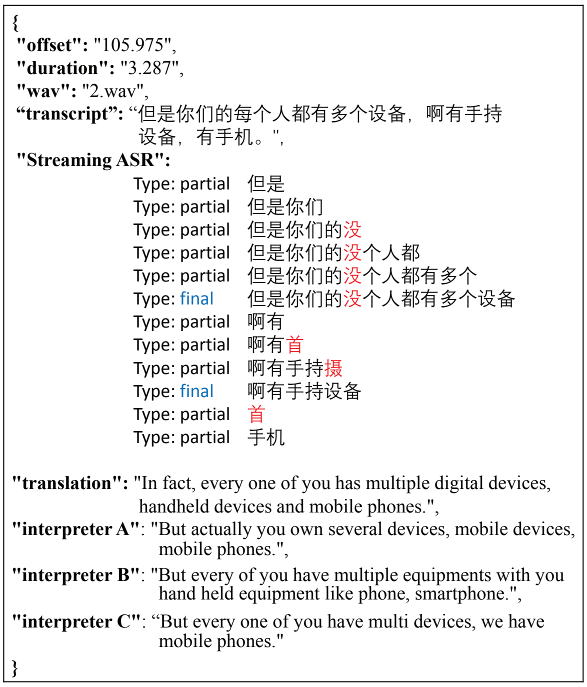
2.4 Human Interpretation

We further ask three experienced interpreters (A, B, and C) with interpreting experience ranging from four to nine years to interpret the six talks of the testset, in a mock conference setting[8].

To evaluate their translation quality, we also ask human translators to evaluate the transcribed interpretation from multiple aspects: adequacy, fluency, and correctness:

- **Rank1**: The translation contains no obvious errors.
- **Rank2**: The translation is comprehensible and adequate, but with minor errors such as incorrect function words and less fluent phrases.
- **Rank3**: The translation is incorrect and unacceptable.

Table 4 shows the translation quality in BLEU and acceptability, which is calculated as the sum of the percentages of Rank1 and Rank2. It shows that their acceptability ranges from 62.8% to 83.0%, but the acceptability and BLEU are not completely positively correlated. This is because human interpreters routinely omit less important information to overcome their limitations in working memory. Acceptability focuses more on accuracy and faithfulness than adequacy, so it can tolerate information omission. Therefore, some information omitted in human interpretation that results in inferior BLEU may not lead to the decrease of acceptability. But BLEU, as a statistical auto-evaluation metric, considers adequacy with the same importance with accuracy. This leads to the discrepancy between BLEU and acceptability.



```
{
 "offset": "105.975",
 "duration": "3.287",
 "wav": "2.wav",
 “transcript”: “但是你们的每个人都有多个设备，啊有手持
                设备，有手机。",
 "Streaming ASR":
                Type: partial   但是
                Type: partial   但是你们
                Type: partial   但是你们的没
                Type: partial   但是你们的没个人都
                Type: partial   但是你们的没个人都有多个
                Type: final     但是你们的没个人都有多个设备
                Type: partial   啊有
                Type: partial   啊有首
                Type: partial   啊有手持摄
                Type: final     啊有手持设备
                Type: partial   首
                Type: partial   手机

"translation": "In fact, every one of you has multiple digital devices,
                handheld devices and mobile phones.",
"interpreter A": "But actually you own several devices, mobile devices,
                  mobile phones.",
"interpreter B": "But every of you have multiple equipments with you
                  hand held equipment like phone, smartphone.",
"interpreter C": “But every one of you have multi devices, we have
                  mobile phones."
}
```

Figure 4: A segment of one example in our test set， including audio, timelines, transcription, translation, streaming ASR results, and interpretation from three human interpreters (only for testing data). The red characters in “Streaming ASR” indicate recognition errors.

Figure 4 lists a segment from one example in our dataset. Notably, we only supply human interpretations for testing data. Here the “Streaming ASR” is the real-time recognition results, in which the “Type:final” means that the audio has detected a pause or silence and thus segmented, and will start to recognize a new sentence, while “Type:partial” is to continue recognizing the current sentence.

3 Experiments

In this section, we introduce our benchmark systems based on the dataset. We conduct experiments on speech translation and simultaneous translation, respectively.

To preprocess the Chinese and the English text, we use an open-source Chinese Segmenter[9], and Moses Tokenizer[10]. After tokenization, we convert

[8] We play the video of the speech, just like in a real simultaneous interpretation scene

[9] https://github.com/fxsjy/jieba

[10] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/

Systems	Test on Transcript		Test on ASR	
	Dev	Test	Dev	Test
pre-train on WMT	20.78	35.13	18.22	33.32
Finetune on <transcript, translation>	**23.47**(2.69↑)	**41.14**(6.01↑)	19.68(1.46↑)	35.71(2.39↑)
Finetune on <ASR, translation>	22.53(1.75↑)	39.23(4.1↑)	**19.82**(1.6↑)	**36.89**(3.57↑)

Table 5: The results of benchmark trained on different training datasets, and evaluated by streaming transcription and ASR input.

all English letters into lower case. To train the MT model, we conduct byte-pair encoding (Sennrich et al., 2016) for both Chinese and English by setting the vocabulary size to 20K and 18K for Chinese and English, respectively. And we use the "multi-bleu.pl" [11] script to evaluate the BLEU score.

3.1 Benchmark System

Our benchmark is a cascade system that includes an ASR module, a sentence segmentation module, and a machine translation (MT) module.

- We use the SMLTA model for ASR, i.e., the streaming transcript/ASR of BSTC is taken as the output of the ASR module.
- The sentence segmentation module is to decide when to translate in real-time. We train a classification model based on the Meaningful Unit (MU) method proposed in Zhang et al. (2020) that implements a 5-class classification (MU, comma, period, question mark, and none). The training data of meaningful units are generated automatically from monolingual sentences based on context-aware translation consistency. The model is pre-trained on ERNIE-base (Sun et al., 2020) and fine-tuned on the transcript of the BSTC training set.
- Once an MU or a sentence boundary (period or question mark) is detected in the sentence segmentation module, the MT module generates translation for the detected sentence. The MT model is firstly pre-trained on the large-scale WMT19 Chinese-English corpus, then fine-tuned on BSTC. The WMT19 corpus includes 9.1 million sentence pairs collected from different sources, *i.e.*, Newswire, United Nations Parallel Corpus, Websites, etc. We use the *big* version of Transformer model in the following experiments.

3.2 Performance of Speech Translation

Speech translation aims at translating accurately without considering system delay. Therefore, we only perform translation when sentence boundaries (periods and question marks) are detected by the sentence segmentation module.

The MT model is firstly trained on WMT, then fine-tuned on 37,901 training pairs of <transcription, translation> and <asr, translation> in two settings, respectively. The purpose of fine-tuning on transcription is to adapt the model to the speech domain, and the purpose of fine-tuning on ASR is to improve the robustness of the MT model against recognition errors. Our model pre-trained on WMT19 achieves a BLEU of 25.1 on Newstest19.

We evaluate our systems on the dev/test set using streaming transcription and streaming ASR as inputs. For each talk in the dev/test set, its streaming text is firstly segmented by the sentence segmentation module, then the translation of each segmentation is concatenated into one long sentence to evaluate the BLEU score. The results are listed in Table 5. Note that the great gap of BLEU in dev and test sets is that, the dev set has only one reference while the testset has 4 references.

Contribution of fine-tuning on speech translation data: The systems pre-trained on WMT obtain an absolute improvement both on clean and noisy input by fine-tuning on <transcription, translation>. The performance of the former model increases by 4.35 BLEU score on average and the latter model obtains 1.93 BLEU score improvement on average. This indicates the transcribed training data can still bring large improvement after pre-training on large-scale training corpus. This probably because it is closer to the test set in terms of the domain (speech) and noise (disfluencies in spoken language).

Contribution of fine-tuning on noisy data: Training on the corpus containing the ASR errors can be effective to improve the robustness of the NMT model. This can be proved by fine-tuning

tokenizer/tokenizer.perl

[11] https://github.com/moses-smt/\mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

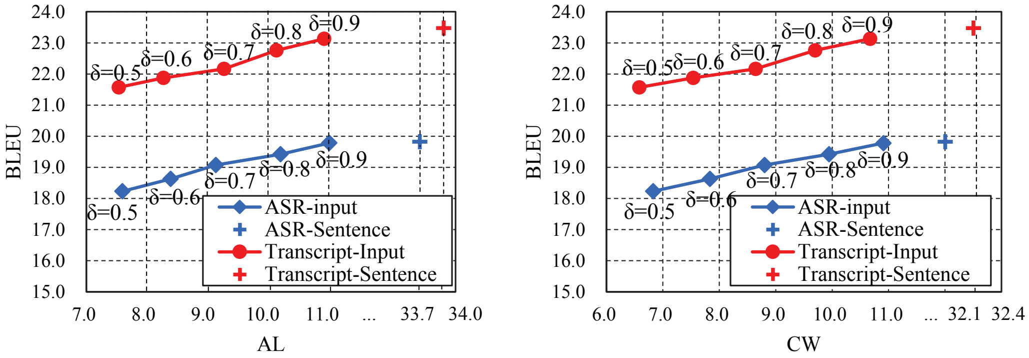


Figure 5: Translation quality against latency metrics on BSTC development set. "ASR-Sentence" and "Transcript-Sentence" denotes the results of full-sentence translation with ASR input and transcript input, respectively.

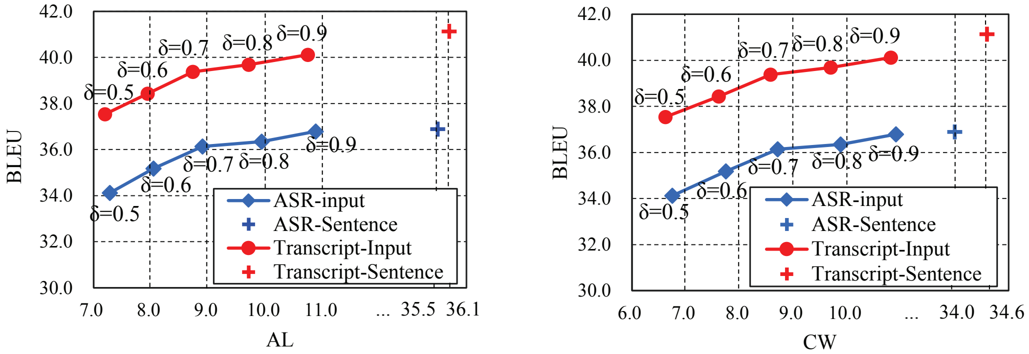


Figure 6: Translation quality against latency metrics on BSTC testset.

on the <ASR, translation> pairs. As shown in the last row of Table 5, the pre-trained model improves 2.93 and 2.59 BLEU scores on average for testing on streaming transcript and streaming ASR, respectively. This manifests that compared with fine-tuning the clean transcription, the model fine-tuned on ASR is less sensitive to false recognition results of ASR.

3.3 Performance of Simultaneous Translation

Different from speech translation, the simultaneous translation should balance translation quality and latency. Therefore, we fix the ASR and MT modules to evaluate our system under different sentence segmentation results. In simultaneous translation, once an MU or a sentence boundary is detected, the MU or sentence is translated immediately. In order to maintain coherent and consistent paragraph translation, we perform context-aware translation following Xiong et al. (2019) that except for the first segment in a sentence, the subsequent segments are translated with force-decoding.

The performance of system on the dev set and test set is listed in Figure 5 and Figure 6, respectively[12]. We use BLEU to evaluate the translation quality and use average lagging (AL) (Ma et al., 2019) and Consecutive Wait (CW) (Gu et al., 2017) as latency metrics. δ is the hyperparameter defined in Zhang et al. (2020) as the thresold of sentence segmentation module. It shows that the translation quality improves consistently with the increase of latency. The AL on both dev and test sets ranges from 7 to 12 and the CW ranges from 6 to 11 for points of simultaneous translation. In addition, we also draw the full-sentence translation results, as denoted by "ASR-Sentence" and "Transcript-Sentences" in the two figures. The full-sentence translation implements a high-latency policy, in which a translation is only triggered when a sentence is received. As shown in the figures, the delay of both "ASR-Sentence" and "Transcript-Sentences" is much higher than the simultaneous translation results.

4 Conclusion and Future Work

In this paper, we release a challenging dataset for the research on Chinese-English speech translation and simultaneous translation. Based on this

[12] We list detailed values in Table 6

	δ	**AL**	**CW**	**BLEU**
Dev Set	**Input ASR**			
	0.5	7.61	6.82	19.07
	0.6	8.42	7.83	19.42
	0.7	9.17	8.80	19.78
	0.8	10.26	9.94	20.25
	0.9	11.08	10.91	20.37
	Input Transcript			
	0.5	7.54	6.58	21.87
	0.6	8.30	7.54	22.16
	0.7	9.31	8.64	22.76
	0.8	10.19	9.70	23.13
	0.9	11.00	10.67	23.62
Test set	**Input ASR**			
	0.5	7.28	6.75	34.12
	0.6	8.04	7.75	35.18
	0.7	8.90	8.71	36.14
	0.8	9.93	9.88	36.35
	0.9	10.87	10.91	36.79
	Input Transcript			
	0.5	7.20	6.62	37.54
	0.6	7.94	7.61	38.43
	0.7	8.73	8.58	39.38
	0.8	9.70	9.70	39.69
	0.9	10.74	10.81	40.12

Table 6: Specific data corresponding to Figure 5 and Figure 6.

dataset, we report a competitive benchmark based on a cascade system. In the future, we will expand this dataset, and propose an effective method to develop an End-to-End speech translation model.

References

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.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Eunah Cho, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2014. A corpus of spontaneous speech in lectures: The KIT lecture corpus for spoken language processing and translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1554–1559, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. 2018. Unsupervised cross-modal alignment of speech and text embedding spaces. *arXiv preprint arXiv:1805.07467*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Christian Federmann and William D Lewis. 2016. Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *International Workshop on Spoken Language Translation*.

Christian Federmann and William D Lewis. 2017. The microsoft speech language translation (mslt) corpus for chinese and japanese: conversational test data for machine translation and speech recognition. *Proceedings of the 16th Machine Translation Summit, Nagoya, Japan*.

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.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation. *Language Resources and Evaluation*.

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.

Jan Niehues, Quan Pham, Thanh Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Interspeech 2018*, pages 1293–1297.

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.

Matthias Paulik and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as mt training data. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 496–501. IEEE.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Collection of a simultaneous translation corpus for comparative analysis. In *LREC*, pages 670–673. Citeseer.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. In *Transactions of the Association for Computational Linguistics*.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 230–238.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Hitomi Tohyama, Shigeki Matsubara, Koichiro Ryu, N Kawaguch, and Yasuyoshi Inagaki. 2004. Ciair simultaneous interpretation corpus. In *Proc. Oriental COCOSDA*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2: A massively multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2007.10310*.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Michael Melese Woldeyohannis, Laurent Besacier, and Million Meshesha. 2017. A corpus for amharic-english speech translation: the case of tourism domain. In *International Conference on Information and Communication Technology for Develoment for Africa*, pages 129–139. Springer.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting. *arXiv preprint arXiv:1907.12984*.

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, Online. Association for Computational Linguistics.