The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022

Weitai Zhang[1,2], Zhongyi Ye[2], Haitao Tang[2], Xiaoxi Li[2], Xinyuan Zhou[2],
Jing Yang[2], Jianwei Cui[1], Pan Deng[1], Mohan Shi[1], Yifan Song[1],
Dan Liu[1,2], Junhua Liu[1,2], and Lirong Dai[1]

[1]University of Science and Technology of China, Hefei, China
[2]iFlytek Research, Hefei, China
{*zwt2021, danliu, jwcui, pdeng, smohan, yfsong*}*@mail.ustc.edu.cn*
lrdai@ustc.edu.cn
{*zyye7, xxli16, httang, xyzhou15, jingyang24, jhliu*}*@iflytek.com*

Abstract

This paper describes USTC-NELSLIP's submissions to the IWSLT 2022 Offline Speech Translation task, including speech translation of talks from English to German, English to Chinese and English to Japanese. We describe both cascaded architectures and end-to-end models which can directly translate source speech into target text. In the cascaded condition, we investigate the effectiveness of different model architectures with robust training and achieve 2.72 BLEU improvements over last year's optimal system on MuST-C English-German test set. In the end-to-end condition, we build models based on Transformer and Conformer architectures, achieving 2.26 BLEU improvements over last year's optimal end-to-end system. The end-to-end system has obtained promising results, but it is still lagging behind our cascaded models.

1 Introduction

This paper describes the submission to IWSLT 2022 Offline Speech Translation task by National Engineering Laboratory for Speech and Language Information Processing (NELSLIP), University of Science and Technology of China, China.

For years, Spoken Language Translation (SLT) has been addressed by cascading an Automatic Speech Recognition (ASR) and a Machine Translation (MT) system. The ASR system processes source speech into source text and the MT system translates ASR output into text in target language independently. Recent trends rely on using a single neural network to directly translate the speech in source language into the text in target language without intermediate symbolic representations. The end-to-end paradigm shows an enormous potential to overcome some of the cascaded systems' problems, such as higher architectural complexity and error propagation (Duong et al., 2016; Berard et al., 2016; Weiss et al., 2017). Last year's results of IWSLT 2021 have confirmed that the performance of end-to-end models is approaching the results of cascaded solutions. The best end-to-end submission (under the same segmentation and training data conditions) is 2 BLEU points (22.6 vs 24.6) below the top-ranked system (Anastasopoulos et al., 2021).

In this work, we build machine translation systems with techniques like back translation (Sennrich et al., 2016a), domain adaptation and model ensemble, which have been proved to be effective practices in IWSLT and WMT (Akhbardeh et al., 2021). Besides, we further improve cascaded speech translation system performance with methods of self-training (Kim and Rush, 2016; Ren et al., 2020; Liu et al., 2019), speech synthesis (Shen et al., 2018), Supervised Hybrid Audio Segmentation (SHAS) (Tsiamas et al., 2022), etc.

In end-to-end condition, we initialize the encoder with the corresponding component of ASR models and the decoder with that of MT models respectively (Le et al., 2021). Methods used in cascaded systems and as much semi-supervised data as possible are utilized to improve end-to-end models' performance. Furthermore, we try to obtain a better performance with ensemble of cascaded and end-to-end models, which may accelerate the application of end-to-end models in industrial scenarios.

The remaining of the paper proceeds as follows. Section 2 describes speech recognition, speech-to-text translation (S2T for short) and text-to-text translation (T2T for short) data used in our experiments. Section 3 and Section 4 present our cascaded and end-to-end systems respectively, where the details about model architectures and techniques for training and inference will be described. The experimental settings and final results are shown in Section 5.

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 198 - 207

2 Datasets and Preprocessing

2.1 Speech Recognition Data

The speech recognition datasets used in our experiments are described in Table 1, in which Librispeech, MuST-C(v1, v2), TED Lium3, Europarl, VoxPopuli and CoVoST are available and used. After extract 40 dimensional log-mel filter bank features computed with a $25ms$ window size and a $10ms$ window shift, we train a baseline ASR model and filter training samples with WER$> 40\%$. Then we augment the speech data with speed perturbation, and over-sample TED/MuST-C corpus with the ratio used last year (Liu et al., 2021), which finally generate almost 8k hours of speech recognition corpora.

Corpus	Duration(h)	Sample Scale
Librispeech	960	1
Europarl	161	1
MuST-C(v1)	399	3
MuST-C(v2)	449	3
TED-LIUM3	452	3
CoVoST2	1985	1
VoxPopuli	1270	1

Table 1: Statistics of ASR Corpora.

We further extend two data augmentation methods: First, Adjacent voices are concatenated to generate longer training speeches; Second, we train a Glow-TTS (Casanova et al., 2021) model with MuST-C datasets and generate 24k hours of audio feature using sentences from EN→DE text translation corpora. The final training data for ASR is described in Table 2.

Data	Duration(h)
Raw data	8276
+ concat	16000
+ oversampling	32000
+ TTS	56000

Table 2: Overall training data for ASR.

2.2 Text Translation Corpora

We participate in translation of English to German, Chinese and Japanese. All available bilingual data and as much monolingual data as possible are used for training our systems. We apply language identification to retain sentences predicted as desired language, remove sentences longer than 250 tokens and with a source/target length ratio exceeding 3, filter sentences with lower scores based on baseline machine translation models. We use LTP4.0[1](Che et al., 2020) for Chinese tokenization, MeCab morphological analyzer[2] for Japanese tokenization and moses for English tokenization. Then subwords are generated via Byte Pair Encoding (BPE) (Sennrich et al., 2016b) with 30k merge operations for each language direction. Table 3 lists statistics of parallel and monolingual data used for training our systems. The details are as follows.

EN→DE The bilingual data includes CommonCrawl, CoVoST2, Europarl, MuST-C(v1, v2), Librivox, News Commentary, Opensubtitles, Parawcrawl(v3, v5.1), Rapid, Wikimatrix-v1 and Wikititles-v2. A total of 151 million sentence pairs are available, 120 million pairs of which are reserved for training. The monolingual English and German data are mainly from News Commentary and News crawl.

EN→ZH Almost 50 million sentence pairs collected from CCMT Corpus, News Commentary, ParaCrawl, Wiki Titles, UN Parallel Corpus, WikiMatrix, Wikititles, MuST-C and CoVoST2 are used for training EN→ZH text MT. 50 million monolingual Chinese sentences are randomly extracted from News crawl and Common Crawl for Back Translation.

EN→JA We use 16 million sentence pairs from MuST-C, CoVoST2, TED Talk, JESC-v2, News Commentary, Paracrawl, Wikimatrix and Wikititles. 20 million Japanese monolingual sentences from News Commentary, News crawl and Common Crawl are randomly extracted for Back Translation.

	Parallel	Monolingual
EN-DE	120M	180M
EN-ZH	50M	50M
EN-JA	15.75M	20M

Table 3: Overall training data for text MT.

2.3 Speech Translation Corpora

The speech translation datasets used in our experiments are described in Table 4. MuST-C and CoVoST2 are available for speech translation (speech, transcription and translation included) in all three

[1] https://github.com/HIT-SCIR/ltp
[2] https://github.com/uenewsar/mecab

language directions, while Europarl is specifically available in EN→DE speech translation track.

We further extend two data augmentation methods: First, transcriptions of all speech recognition datasets are sent to a text translation model to generate text y' in target language, which is similar with sentence knowledge distillation. The generated y' with its corresponding speech are directly added to speech translation dataset (described as KD Corpus in Table 4). Second, we use the trained Glow-TTS model to generate audio feature from randomly selected sentence pairs from EN→DE, EN→ZH and EN→JA text translation corpora. The generated filter bank features and their corresponding target language text are used to expand our speech translation dataset (described as TTS Corpus in Table 4).

	Corpus	Duration(h)	Sample Scale
EN-DE	Europarl	161	2
	MuST-C	449	2
	CoVoST2	1094	2
	KD	16000	2
	TTS	24000	1
EN-ZH	MuST-C	593	2
	CoVoST2	1092	2
	KD	16000	2
	TTS	27000	1
EN-JA	MuST-C	282	2
	CoVoST2	988	2
	KD	16000	2
	TTS	13000	1

Table 4: Statistics of Speech Translation Corpora

3 Cascaded Speech Translation

3.1 Automatic Speech Recognition

Voice Activity Detection We use Supervised Hybrid Audio Segmentation (SHAS) (Tsiamas et al., 2022) to split long audios into shorter segments. SHAS is originally propsed to learn the optimal segmentation for speech translation. Experiments on MuST-C and mTEDx show that the translation of the segments produced by SHAS approaches the quality of the manual segmentation on 5 languages pairs. Hence, we use SHAS for both Voice Activity Detection in ASR and segmentation in Speech Translation, which means that we have no more segmentation operations and ASR outputs are directly sent to text machine translation component.

Besides, we propose a semantic VAD method as follows: 1) train a text segmentation model based on transformer; 2) re-segment ASR results into new sentences with complete semantic information; 3) use Force Alignment to align speech time stamp and ASR results; 4) re-segment voices into new fragments. We hope to seek a more friendly segmentation for machine translation.

Model Architecture We think representations sent to ASR encoder component are important, so we use three model architectures in ASR: VGG-Transformer (Mohamed et al., 2019), VGG-Conformer (Gulati et al., 2020) and GateCNN-Transformer (Dauphin et al., 2017) implemented on Fairseq, described as follows:

- VGG-Conformer: 2 layers of VGG and 12 layers of Conformer in encoder, 6 layers of Transformer in decoder. The embedding size is 512, the hidden size of FFN is 2048, and the attention head is 8.
- VGG-Transformer: 2 layers of VGG and 16 layers of Transformer in encoder, 6 layers of Transformer in decoder. The embedding size is 512, the hidden size of FFN is 4096, and the attention head is 8.
- GateCNN-Conformer: 6 layers of GateCNN and 12 layers of Conformer in encoder, 6 layers of Transformer in decoder. The embedding size is 512, the hidden size of FFN is 2048, and the attention head is 8.

The Specaugment technique (Park et al., 2019) is used to improve robustness, and Connectionist Temporal Classification (CTC) is added to make models converge better. Other training details are as follows: 1) We apply BPE to the transcripts with 30000 merge operations; 2) Arabic numerals are converted into corresponding English words; 3) Punctuation marks and uppercase are remained for fitting text machine translation; 4) We use Adam optimizer and adopt the default learning schedule in fairseq; 5) Model is trained on 32 Tesla V100 40G GPUs within 2 days; 6) We use ensemble decoding of several models with beamsize of 15 to produce final transcriptions; 7) Other parameters are default in Fairseq.

3.2 Neural Machine Translation

The machine translation models are based on Transformer (Vaswani et al., 2017) implemented on the Fairseq toolkit (Ott et al., 2019). Each single model is carried out on 16 NVIDIA V100 GPUs with default settings. Important techniques used in our experiments are: Back Translation, Sentence-level Knowledge Distillation, Domain Adaptation and Ensemble.

Back Translation Back-Translation (Sennrich et al., 2016a) is an effective way to improve the translation performance by translating target-side monolingual data to generate synthetic sentence pairs, which has been widely used in research and industrial scenarios. We train NMT models with bilingual data, and translate German/Chinese/Japanese sentences to English ones.

Knowledge Distillation Sentence-level Knowledge Distillation (Kim and Rush, 2016)(also known as self-training) is another useful technique to improve performance. We augment training data by translating English sentences to German/Chinese/Japanese using a trained NMT model.

Domain Adapatation As high-quality and domain-specific translation is crucial, fine-tuning the concatenation system on in-domain data shows the best performance (Saunders, 2021). To improve in-domain translation while do not decrease the quality of out-domain translation, we fine-tune the NMT model on a mix of in-domain data (MuST-C, TED-LIUM3, etc.) and random selected out-of-domain data until convergence. The speech recognition training data are also used as augmented in-domain self-training data by translating the labelled English sentences.

We also use Denoise-based approach (Wang et al., 2018) to measure and select data for domain MT and apply them to denoising NMT training. Denoising is concerned with a different type of data quality and tries to reduce the negative impact of data noise on MT training, in particular, neural MT (NMT) training.

Ensemble For each language direction, we train 4 variants based on Transformer big settings and the final model is the ensemble of the 4 models:

- E12D6: 12 layers for the encoder and 6 layers for the decoder. The embedding size is 1024, FFN size is 8192 and attention head is 16. All available corpora including bilingual, BT and FT data are used.
- E15D6: 15 layers for the encoder, 10% training data are randomly dropped and a different seed is set.
- E18D6: 18 layers for the encoder and 10-30% training data with lower machine translation scores are dropped.
- Macaron: A model with macaron architecture (Lu et al., 2019) based on data of E18D6. 36 layers for the encoder and FFN size is 2048.

3.3 Robust MT Training

To address the error propagation problem in cascaded ST, we propose a ASR output adaptation training method for improving MT model robustness against ASR errors. English transcriptions of all speech translation datasets are sent to a trained ASR model to generate text $x^{'}$ in source side, paired with the target side labels. We use 3 approaches to improve MT model's robustness detailed as follows: 1) We use the synthetic data to fine-tune MT model; 2) While fine-tuning, we add KL loss to prevent over-fitting; 3) we distill the model both by clean input and ASR output as showed in Figure 1.

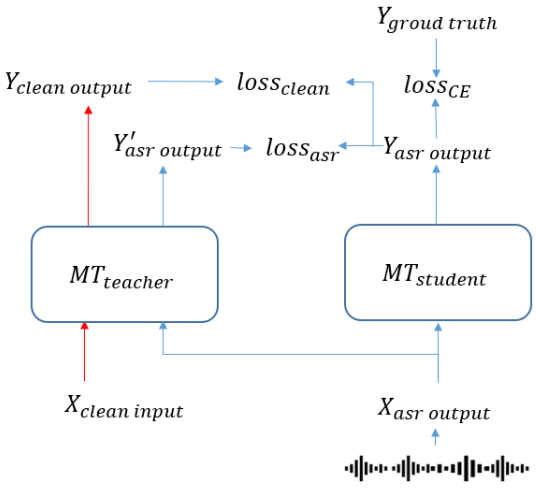

Figure 1: Overview of Robust MT Training.

4 End-to-End Speech Translation

As regards model architecture, we investigate 4 variants in end-to-end speech translation.

- VGG-C: The encoder is VGG-Conformer, initialized by ASR VGG-Conformer architecture.

The decoder is 6 layers of Transformer with embedding size of 1024, attention head of 16 and FFN size of 8192.

- VGG-C-init: The encoder is VGG-Conformer, initialized by ASR VGG-Conformer architecture. The decoder is 6 layers of Transformer, initialized by NMT E15D6 variant.
- VGG-T: The encoder is VGG-Transformer, initialized by ASR VGG-Transformer architecture. The decoder is 6 layers of Transformer with embedding size of 1024, attention head of 16 and FFN size of 8192.
- VGG-T-init: The encoder is VGG-Transformer, initialized by ASR VGG-Transformer architecture. The decoder is 6 layers of Transformer, initialized by NMT E15D6 variant.

5 Experiments

All our experiments are conducted using Fairseq toolkit (Ott et al., 2019). We use word error rate (WER) to evaluate the ASR models and report case-sensitive SacreBLEU scores for machine translation. Results of MuST-C tst-COMMON (tst-COM), IWSLT tst2018/tst2019/tst2020 are listed together, which can be compared as baselines for other researchers and participants in the future. We also present results of IWSLT 2022 testsets in the Appendix.

5.1 Automatic Speech Recognition

The overall experimental results about ASR is described in Table 6. We use SHAS as a segmentation tool in default for all testsets. We compare the results of different model architectures with and without TTS augmented training data, showed in line 1-6. In our experiments, TTS augmented data has consistent improvements in all three architectures, and an absolute gain of 0.42% accuracy is observed in GateCNN-Conformer, which makes GateCNN-Conformer with TTS augmented data performs best as a single model.

In line 7, we ensemble all 6 single models to gain a best result, where the WER is at an average of 5.32, and 0.69 lower than the best single model. For comparison with other works, we list the result of tst-COM with official segments in line 8, which performs better than concatenating the segments and using SHAS. In line 9, we present results with semantic SHAS (described in Sec. 3.1) based on the ensemble models, which shows that semantic SHAS is slightly worse and lagging behind SHAS by 0.13 in accuracy. In our final submissions, line 7 serves as the ASR part of our cascaded primary system, and line 9 serves as part of a contrastive system.

5.2 Speech Translation

For text machine translation, we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. To speed up the training process, we conduct training with half precision floating point (FP16). We set max learning rate to $7e$-4 and warmup-steps to 8000. To improve model robustness, we set bpe dropout to 0.05, and mask 15% words in source and target inputs in accord with BERT. When fine tuning on in-domain datasets, we add KL loss with weight=1.0 to avoid over-fitting.

For end-to-end ST, the segmentation tool used is SHAS (to our knowledge, using semantic SHAS will not be considered as end-to-end). All available training data including TTS augmented data and knowledge distillation data described in Sec. 2.3 are used. We also fine-tune models on in-domain corpus for further improvements.

For tst-COM, we report results of both official segmentation and SHAS segmentation. Sacrebleu scores are computed by using automatic resegmentation of the hypothesis based on the reference translation by mwerSegmenter.

Effectiveness of Robust MT Training The experiment is conducted based on EN→DE cascaded speech translation track. We generate 1.38M sentences from 1500h speech translation datasets. Experimental results are described in Table 5. By comparing line 3 and line 6, our method can further gain 0.55 and 0.75 BLEU in tst-COM and tst2018 regardless of the impact of domain adaptation. Robust MT Training is adopted for training all our following systems.

EN→DE Experimental results are described in Table 7. In the first group of text MT results, line 2-5 show the effectiveness of model size, data clean and fine-tuning on in-domain datasets. We ensemble 4 different variants described in Sec. 3.2 and constitute results in line 6, which makes our text MT outperforming Volctrans's ensemble results (Zhao et al., 2021) by 1.85 BLEU in tst-COM.

In the second group of cascaded ST results, we present final results produced with ensemble ASR

#		tst-COM	tst2018
1	text MT	36.21	32.14
2	ASR→text MT	33.34	26.20
3	+finetune	34.11	28.41
4	Robust Training	34.21	27.62
5	+KL Loss	34.61	28.69
6	+KD Loss	34.66	29.16

Table 5: Experimental results of Robust MT Training.

in Table 6 and ensemble text MT in line 6 by SHAS and semantic SHAS respectively. By comparing line 8 and line 9, SHAS performs better in tst-COM and tst2018, while semantic SHAS performs better in tst2019 and tst2020. Results of tst-COM for speech translation contained in parenthesis are based on official segmentation, which means our cascaded system outperforms Volctrans's cascaded results (Zhao et al., 2021) by 2.72 BLEU in tst-COM. We observe more improvements in cascaded ST than text MT due to our better ASR system.

Regards end-to-end ST, we compare the results of different model architectures with and without TTS augmented training data, showed in line 11-16. From line 11-14, TTS augmented data has improvements by 0.43 BLEU in VGG-Conformer-init, while decrease the BLEU scores (0.09) in VGG-Conformer. Using NMT decoder for initialization brings consistent improvements with or without TTS data. In line 17, we ensemble all 6 single models with outperforming best single model by an average of 0.97 BLEU, but it is still lagging behind cascaded systems by 1.36 BLEU in tst-COM. Our end-to-end system outperforms KIT's end-to-end results (Nguyen et al., 2021) by 2.26 BLEU in tst-COM.

To investigate the effectiveness of ensemble of cascaded and end-to-end systems, we present the results in line 18 and 19 with SHAS and semantic SHAS respectively. We observe consistent and slight improvements in all testsets except tst-COM using SHAS. We submit systems of #8, #9, #17, #18, #19, with #8 as primary system in cascaded condition and #17 as primary system in end-to-end condition.

EN→ZH Experimental results are described in Table 8. Regards text MT, line 1-3 show the effectiveness of model size and data clean. We further improve performance with fine-tuning models on MuST-C and TED Talk corpus in line 4. Line 5 shows results of ensemble MT from 4 fine-tuned variants described in Sec. 3.2. In the second group of cascaded ST results, we present final results produced with ensemble ASR in Table 6 and ensemble text MT by SHAS and semantic SHAS respectively. Results of tst-COM for speech translation contained in parenthesis are based on official segmentation. Regards end-to-end ST, we train 4 different models based on conclusions from EN→DE end-to-end experiments. In line 12, we ensemble 4 single models and get 28.92 BLEU in tst-COM within official segmentation. Our final end-to-end ST result on tst-COM is still lagging behind cascaded system by 0.89 BLEU.

Same with EN→DE translation track, we present the ensemble results of cascaded systems and end-to-end systems in line 13 and 14 with SHAS and semantic SHAS respectively, which brings slight improvements comparing with cascaded system. We submit systems of #6, #7, #12, #13, #14, with #6 as primary system in cascaded condition and #12 as primary system in end-to-end condition.

EN→JA The overall experimental results is described in Table 9. Regards text MT, line 1-3 show the effectiveness of model size and data clean. We further improve performance with fine-tuning models on MuST-C and TED Talk corpus in line 4. Line 5 shows results of ensemble models from 4 fine-tuned variants described in Sec. 3.2. Line 6-7 present cascaded ST results with ASR outputs from ensemble models, which only decrease 0.25 BLEU on dev and 0.48 BLEU on tst-COM compared with text MT. The reason might be partly attributed to the fact that text MT BLEU is relatively lower and ASR errors have a smaller portion of all factors affecting the performance. While MuST-C training data and tst-COM have no punctuations in Japanese side, We think punctuations help people understand. We train a punctuation model based on transformer encoder, and add punctuations for translations. The performance decreases because of the mismatch between references and translations in punctuations.

Regards end-to-end ST, we train 4 different models based on conclusions from EN→DE end-to-end experiments. In line 12, we ensemble 4 models and get 18.61 BLEU in tst-COM with official segmentation. Our final end-to-end ST result on tst-COM is still lagging behind cascaded system by 2.89 BLEU. We submit systems of #6, #7, #8, #9, #12, with #6 as primary system in cascaded condition and #12 as primary system in end-to-end condition.

#	System	tst-COM	tst2018	tst2019	tst2020	avg
1	VGG-Conformer (w/ TTS)	3.66	8.56	5.28	7.23	6.18
2	VGG-Conformer (w/o TTS)	3.70	8.55	5.34	7.54	6.28
3	VGG-Transformer (w/ TTS)	**3.31**	8.39	5.58	7.43	6.18
4	VGG-Transformer (w/o TTS)	3.34	8.50	5.85	7.76	6.36
5	GateCNN-Conformer (w/ TTS)	4.06	7.87	5.14	6.98	6.01
6	GateCNN-Conformer (w/o TTS)	4.35	8.12	5.74	7.52	6.43
7	ensemble (1, 2, 3, 4, 5, 6, SHAS)	3.36	**7.30**	**4.59**	**6.03**	**5.32**
8	7 (w/o SHAS)	3.49	-	-	-	-
9	7 (w/ semantic SHAS)	3.54	7.26	4.89	6.10	5.45

Table 6: Overall experimental results of ASR. We present WER performance of tst-COM, tst2018, tst2019 and tst2020, and hope it can be compared as baselines in other works. For tst-COM, we concatenate the audios and segment with SHAS except for line 8.

#	Systems	tst-COM	tst2018	tst2019	tst2020
	Text MT				
1	Volctrans(ensemble) (Zhao et al., 2021)	(36.7)	-	-	-
2	base	32.65	29.02	26.90	-
3	clean+big	36.21	32.03	29.64	-
4	text MT	36.84	32.65	30.02	-
5	4+finetune	38.20	34.56	**31.86**	35.54
6	ensemble MT	**38.55**	**34.89**	31.82	**36.08**
	Cascaded ASR→MT				
7	Volctrans(ensemble) (Zhao et al., 2021)	(33.3)	-	-	-
8	ensemble ASR→6+SHAS	**34.73(36.02)**	**30.02**	29.25	32.15
9	+semantic SHAS	34.36*(36.02)	29.59	**29.40**	**32.44**
	End-to-End ST				
10	KIT (ensemble) (Nguyen et al., 2021)	(32.4)	-	-	-
11	VGG-C (w/o TTS)	31.81(33.37)	28.47	26.48	28.82
12	VGG-C-init (w/o TTS)	31.79(33.48)	28.44	26.70	29.17
13	VGG-C (w/ TTS)	31.58(32.78)	**29.00**	26.47	28.69
14	VGG-C-init (w/ TTS)	**32.39(33.74)**	28.98	**27.03**	**29.59**
15	VGG-T (w/ TTS)	31.37(32.72)	28.54	26.17	28.42
16	VGG-T-init (w/ TTS)	31.21(32.81)	28.68	26.23	28.67
17	Ensemble (11-16)	**33.23(34.66)**	**29.93**	**28.20**	**30.57**
	Ensemble of cascaded and e2e systems				
18	Ensemble(8, 17)	33.58(36.05)	**30.93**	**29.57**	32.15
19	Ensemble(8, 17)* +semantic SHAS	**34.47*(36.13)**	30.19	29.41	**32.50**

Table 7: Overall experimental results of EN→DE translation track. Results of tst-COM for speech translation contained in parenthesis are based on official segmentation which are comparable with previous works. Results with * are based on semantic SHAS, and others are based on SHAS. Weights of models in line 18 and 19 are different. We submitted 5 systems in EN→DE track with system ID in bold.

#	Systems	tst-COM
	Text MT	
1	base	23.26
2	clean+big	26.92
3	text MT	27.49
4	3+finetune	30.19
5	ensemble MT	**31.03**
	Cascaded ASR→MT	
6	ensemble ASR→5+SHAS	**29.68(29.81)**
7	+semantic SHAS	29.23(29.81)
	End-to-End ST	
8	VGG-C (w/ TTS)	28.34(28.60)
9	VGG-C-init (w/ TTS)	28.51(28.71)
10	VGG-T (w/ TTS)	27.91(28.41)
11	VGG-T-init (w/ TTS)	27.85(28.23)
12	Ensemble (8,9,10,11)	**28.78(28.92)**
	Ensemble of cascaded and e2e systems	
13	Ensemble(6, 12)	29.80(29.79)
14	+semantic SHAS	29.41(29.79)

Table 8: Overall experimental results of EN→ZH translation track. Results in parentheses are with official segmentation.

#	Systems	tst-COM
	Text MT	
1	base	15.44
2	clean+big	17.43
3	text MT	18.72
4	3+finetune	21.78
5	ensemble MT	**22.02**
	Cascaded ASR→MT	
6	ensemble ASR→5+SHAS	**21.25(21.50)**
7	+semantic SHAS	21.11(21.50)
8	6+punctuation model	19.29(18.81)
9	7+punctuation model	19.84(18.81)
	End-to-End ST	
8	VGG-C (w/o TTS)	17.72(17.71)
9	VGG-C-init (w/o TTS)	17.66(17.76)
10	VGG-C-init (w/ TTS)	**17.97(18.20)**
11	VGG-T-init (w/ TTS)	17.60(17.66)
12	Ensemble (8,9,10,11)	**18.62(18.61)**

Table 9: Overall experimental results of EN→JA translation track. Results in parentheses are with official segmentation.

6 Conclusion

This paper summarizes the results of IWSLT 2022 Offline Speech Translation task produced by the USTC-NELSLIP team. We investigate various model architectures and data augmentation approaches to build strong speech translation systems, both in cascaded condition and end-to-end condition. In our experiments, we demonstrate the effectiveness of Back Translation, Knowledge Distillation, Domain Adaptation, Ensemble, elegant segmentation. Our end-to-end model surpasses the last year's best system by 2.26 BLEU, but it is still lagging behind our cascaded model by an average of 1.73 BLEU scores on MuST-C test sets. As a note for future work, we would like to investigate the effectiveness of speech data augmentation and multi-modal representations in end-to-end speech translation.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. 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.

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

Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico San-

tos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluísio, and Moacir Antonelli Ponti. 2021. Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3645–3649. ISCA.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: A open-source neural chinese language technology platform with pretrained models. *arXiv preprint arXiv:2009.11616*.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941. PMLR.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

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

Hang Le, Florentin Barbier, Ha Nguyen, Natalia Tomashenko, Salima Mdhaffar, Souhir Gabiche Gahbiche, Benjamin Lecouteux, Didier Schwab, and Yannick Estève. 2021. ON-TRAC' systems for the IWSLT 2021 low-resource speech translation and multilingual speech translation shared tasks. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 169–174, Bangkok, Thailand (online). Association for Computational Linguistics.

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021. The USTC-NELSLIP systems for simultaneous speech translation task at IWSLT 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation, IWSLT 2021, Bangkok, Thailand (online), August 5-6, 2021*, pages 30–38. Association for Computational Linguistics.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Understanding and improving transformer from a multi-particle dynamic system point of view. *CoRR*, abs/1906.02762.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. Transformers with convolutional context for ASR. *CoRR*, abs/1904.11660.

Tuan Nam Nguyen, Thai Son Nguyen, Christian Huber, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, and Sebastian Stüker. 2021. KIT's IWSLT 2021 offline speech translation system. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 125–130, Bangkok, Thailand (online). Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

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, Online. Association for Computational Linguistics.

Danielle Saunders. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR*, abs/2104.06951.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta Ruiz Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *ArXiv*, abs/2202.04774.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 133–143. Association for Computational Linguistics.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581.

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. In *Proceedings of the 18th International Conference on Spoken Language Translation, IWSLT 2021, Bangkok, Thailand (online), August 5-6, 2021*, pages 64–74. Association for Computational Linguistics.

A Appendix

We present results of official test sets and progress test sets. For En→DE translation track, end-to-end model is lagging behind cascaded model by 1.4 BLEU on tst2022 and 1.8 BLEU on tst2021. Our best result surpasses the last year's best system by 4.4 BLEU, which means performant systems built with classical approaches are strongly competitive. In English to Japanese track, results with punctuations added performs better in ref2 and worse in ref1, mostly because of reference annotations.

#	ref2	ref1	both
8	26.7	23.9	37.6
9	26.3	23.7	37.1
17	25.3	22.9	35.7
18	26.6	23.8	37.4
19	26.2	23.7	37.0

Table 10: Offical BLEU results of IWSLT tst2022 in EN→DE speech translation track.

#	ref2	ref1	both
HW-TSC	24.6	20.3	34.0
8	28.9	24.1	40.3
9	29.0	23.8	40.1
17	27.2	23.0	38.4
18	29.0	23.9	40.3
19	28.8	23.7	39.8

Table 11: Offical BLEU results of IWSLT tst2021 in EN→DE speech translation track.

#	ref2	ref1	both
6	35.8	35.7	44.1
7	35.5	35.3	43.7
12	33.8	34.1	41.9
13	36.1	36.0	44.5
14	35.7	35.5	44.0

Table 12: Offical BLEU results of IWSLT tst2021 in EN→ZH speech translation track.

#	ref2	ref1	both
6	21.6	20.1	33.4
7	21.2	19.8	32.8
8	24.9	18.3	35.2
9	23.8	18.4	34.3
12	20.5	17.4	30.5

Table 13: Offical BLEU results of IWSLT tst2021 in EN→JA speech translation track.