Model	En-De		En-Fr		En-Zh		En-Ja	
	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)	WER(↓)	BLEU(↑)
Text MT	/	22.19	/	30.68	/	25.01	/	22.93
Pipeline	16.19	19.50	14.20	26.62	14.20	21.52	14.21	**20.87**
E2E	16.19	16.07	14.20	27.63	14.20	19.15	14.21	16.59
Multi-task	15.20	18.08	13.04	28.71	13.43	20.60	14.01	18.73
Two-stage	15.18	19.08	13.34	**30.08**	13.55	20.99	14.12	19.32
Interactive	**14.76**	**19.82**$^{*}_{\star}{}^{\ddagger}$	**12.87**$^{*}_{\star}{}^{\ddagger}$	29.79$^{*}_{\star}$	**13.38***	**21.68**$_{\star}{}^{\ddagger}$	**13.91**$^{*}_{\star}{}^{\ddagger}$	19.60$_{\star}{}^{\ddagger}$

Table 1: Evaluation of speech recognition and speech translation on TED En-De/Fr/Zh/Ja datasets. E2E denotes to the pretrained end-to-end ST model, and Interactive represents our proposed interactive learning model. $*$, $\star$, and $\ddagger$ indicate Interactive learning model is statistically significant ($p < 0.01$) compared with Pipeline, Multi-task, and Two-stage, respectively.

BLEU scores in the first row are the translation results by text MT model when the clean manual transcriptions are given as inputs. This can be seen as the upper bound for speech translation task. We set $\lambda = 0.3$ and $k = 3$ in the interactive learning model.

Similar Languages We first analyze En-De and En-Fr language pairs. From the first two rows, we can see that the translation quality drops dramatically when the output of ASR model is fed as the input to the MT model compared with the clean transcriptions input. It indicates that text MT model is very sensitive to recognition errors, which is one of the main problems in the pipeline system. Pretrained end-to-end ST model outperforms the pipeline system by 0.99 BLEU points on En-Fr language direction, but it does not show superiority on En-De. We argue that end-to-end model may have superiority of less error propagation on more similar language pairs, such as En-Fr or En-Es. This is consistent with **?** (**?**) who conducted experiments on En-Es and found end-to-end ST has better performance than the pipeline system. Compared with the end-to-end model, multi-task learning model can obtain some improvements, which improves 2.01 and 0.98 BLEU scores for En-De and En-Fr, respectively. However, with information exchanging, our proposed interactive learning model significantly outperforms multi-task learning model on the quality of both speech recognition and speech translation. It demonstrates the effectiveness of the interactive attention mechanism. Although our method does not outperform two-stage model on En-Fr speech translation task, it has a better performance on ASR result. The underlying reason is that the goal of two-stage model is to optimize the translation quality with the information of complete transcription while ignoring the recognition, so it can improve the translation quality but leave the recognition alone.

Dissimilar Languages It is even more difficult to implement end-to-end speech translation on dissimilar language pairs, such as En-Zh and En-Ja. Because these kind of models are required to learn not only the alignments between source frames and translation words, but also the word orders in long distances. Therefore, in our experiments, most of the end-to-end models are inferior than pipeline system. However, the proposed interactive learning model can significantly outperform end-to-end ST model, traditional multi-task learning model and two-stage model, approaching to or slightly better than pipeline system.

λ	Dev		Test	
	WER	BLEU	WER	BLEU
0.0	14.87	15.74	13.43	20.60
0.1	**14.47**	15.93	**12.92**	20.88
0.3	14.51	**16.28**	13.24	**21.01**
0.5	15.50	15.66	14.17	20.68
1.0	15.92	15.06	14.52	20.13

Table 2: The performance of speech recognition and speech translation under different hyper-parameters λ on the En-Zh *Dev* set and *Test* set.

Wait-k	Dev		Test	
	WER	BLEU	WER	BLEU
Wait-0	14.51	16.28	13.24	21.01
Wait-1	14.29	16.09	**13.17**	21.30
Wait-3	**14.24**	**16.74**	13.38	**21.68**
Wait-5	14.36	16.55	13.51	21.45

Table 3: The performance of speech recognition and speech translation with different word latency in wait-k policy on the En-Zh *Dev* set and *Test* set.

5.5 Effect of the Hyper-parameters

We investigate how much information from two tasks should be taken into consideration in the interactive attention sublayer. Table 2 reports the WER and BLEU scores under different λ on the En-Zh. If $\lambda = 0.0$, the model degrades to traditional multi-task learning model which does not utilize any information from the other task. As shown in the table, as λ increases, both recognition quality and translation quality can be improved with information interacting. When $\lambda = 0.3$, our interactive learning model achieves the best performance on the speech translation task. However, λ can not be too large, otherwise two tasks may interfere with each other and affect its own performance. Therefore, we use $\lambda = 0.3$ for all experiments.

5.6 Effect of k in Wait-k Policy

We then investigate the effect of word latency in wait-k policy on En-Zh language pairs. As shown in Table 3, the speech translation quality in BLEU scores can be improved with the increase of word latency. It indicates that the speech

Model	Params	Speed	
		Train	Inference
Pipeline	122.4M	/	10.89
E2E	61.2M	4.73	16.17
Multi-task	61.2M	4.41	16.26
Two-stage	92.7M	1.13	7.44
Interactive	61.2M	4.23	11.98

Table 4: Statistics of parameters, training and inference speeds. The number in Train denotes the average number of training steps per second. The number in Inference is the average amount of sentences generated per second.

translation task can become easier if more source information from the same modality is given. However, as k increases, it will affect the performance of speech recognition task. If $k \to \infty$, this model degrades to the analogous two-stage model. Then the speech translation task can obtain the information from complete transcribed sentence, while speech recognition task can not utilize any information from translations. The interactive learning model has the best performance when $k = 3$.

5.7 Parameters and Speeds

The parameter sizes of different models are shown in Table 4. The pipeline system needs a separate ASR model and MT model, so its parameters are doubled. Two-stage model has 1.5 times larger parameters since it has two different decoders in two stages. In multi-task learning model and interactive learning model, we share the parameter between different tasks. Therefore, they have the same number of parameters with end-to-end model. Table 4 also shows the training and inference speed of different models on En-Zh test set. The training speed of interactive learning model is 4.23 steps per second, which is comparable with the end-to-end model but is much faster than two-stage model. During inference, the average decoding speed of interactive learning model is 11.98 utterances per second. Although it is slower than end-to-end model and multi-task learning model, it can generate transcriptions paired with translations in one model synchronously. While two-stage model can also generate transcription and translation in a single model, its implementation which is in a cascade manner is much slower even than pipeline system.

5.8 Case Study

We show the case study in Figure 4. In pipeline system, ASR model first recognizes the speech utterance into "brainstormed on solutions to the best child is facing their city". Since it wrongly recognizes "the biggest challenges" into "the best child is", text MT then translates the incorrect recognition phrase, resulting the result is far from the reference. It is more difficult for the end-to-end ST model to generate a correct translation and its output is totally wrong. This model may comprehend the speech of "brainstorm" into "buhrstone" which has a similar pronunciation and it omits the translation of "the biggest". Although the multi-task learning model has an enhanced acoustic encoder, it repeatedly attends to the speech of "storm" without transcription as guidance and translates it twice. As for two-stage model, it erroneously recognized "the biggest" into "the best" in the first stage based on which the second decoder also gives a wrong translation. Compared to the above approaches, our model generates the right transcription and translation through interactive attention mechanism, which matches the reference best.

Reference	brainstorm on solutions to the biggest challenges facing their city
	集思广益 想 办法 解决 城市 面临 的 最大 难题
Pipeline	对 最好 的 孩子们 (the best child is) 实施 头脑 风暴 他们 要 面对 他们 的 城市
E2E	带 着 石灰岩 (buhrstone) 的 解决方案，带 着 他们 的 城市 面临 挑战
Multi-task	头脑 风暴 风暴 (storm)，解决 城市 面临 的 最大 的 挑战
Two-stage	头脑 风暴 解决 了 城市 面临 的 最好 的 (the best) 挑战
Interactive	头脑 风暴 解决 了 城市 面临 的 最大 挑战

Figure 4: An Example of speech translation generated by different models. Words in blue and green are original words in the manual transcription, corresponding translation reference and correct translations with the similar meaning, while words in red are the wrong translations.

6 Conclusion and Future Work

In this paper, we propose an interactive learning model to conduct speech recognition and speech translation interactively and simultaneously. The generation process of recognition and translation in this model can not only utilize the already generated outputs, but also the outputs generated in the other task. We then present a wait-k policy which can further improve the speech translation quality. Experimental results on different language pairs demonstrate the effectiveness of our model. In the future, we plan to design a streaming encoder and make a step forward in achieving end-to-end simultaneous interpretation.

7 Acknowledgments

The research work described in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2016QY02D0303, the Natural Science Foundation of China under Grant No. U1836221 and 61673380, and Beijing Municipal Science and Technology Project No. Z181100008918017 as well. The research work in this paper has also been supported by Beijing Advanced Innovation Center for Language Resources.

Rem illum minus mollitia pariatur maiores ipsam, doloribus mollitia nihil porro autem minus perferendis repellendus molestias cupiditate expedita fugit, sapiente tempora autem sit perferendis harum inventore, eius rerum molestiae minima aliquam dicta natus officiis est quia earum?Tempora nostrum exercitationem, nulla maxime cum,