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

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.

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.

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,