SimulS2S: End-to-End Simultaneous Speech to Speech Translation

Anonymous authors
Paper under double-blind review

Abstract

Simultaneous speech to speech translation aims to interpret concurrently with the speech in source language, which is of great importance to the real-time understanding of spoken lectures or conversations. Previous methods usually divide this problem into three stages: simultaneous automatic speech recognition (ASR), simultaneous neural machine translation (NMT), and simultaneous text to speech (TTS), which is not end-to-end and suffers from translation delay and error propagation. In this work, we propose SimulS2S, an end-to-end simultaneous speech to speech translation system that directly translates from source-language speech into target-language speech concurrently, which jointly optimizes speech recognition, text translation and speech synthesis in one sequence to sequence model. SimulS2S consists of a speech encoder and a speech decoder both with a speech segmenter and a wait-k strategy for simultaneous translation. Since simultaneous speech to speech translation is challenging, we propose several key techniques to help the training of SimulS2S: 1) a curriculum learning mechanism to train the model gradually from full-sentence translation to simultaneous translation; 2) two auxiliary tasks: ASR and S2T (speech to text translation) that share the same encoder with SimulS2S model to help the training of the encoder; 3) knowledge distillation to transfer the knowledge from the cascaded NMT and TTS models to the SimulS2S model. Experiments on Fisher Spanish-English conversation translation datasets demonstrate that SimulS2S 1) achieves low translation delay and reasonable translation quality compared with full-sentence speech to speech translation (without simultaneous translation), and 2) although performs worse than but close to the accuracy of simultaneous translation with three-stage cascaded models, demonstrating the potential of end-to-end approach for this challenging task.

1 Introduction

Simultaneous speech to speech translation (Fügen et al., 2007; Bangalore et al., 2012; Oda et al., 2014; Sarkar, 2016) translates source-language speech into target-language speech concurrently, which plays an important role to narrow the language barrier and ensure real-time understanding of spoken lectures or conversations, and is now widely used in many scenarios such as international conferences in multilateral organizations (UN/EU). The two key requirements for simultaneous speech translation are translation accuracy and translation delay (Mieno et al., 2015), which make it extremely challenging compared with full-sentence translation.

Previous works on simultaneous speech to speech translation (Bangalore et al., 2012; Sarkar, 2016) employ a cascaded approach with three stages: simultaneous automatic speech recognition (ASR) (Rao et al., 2017) that transcribes the source speech into source text in real time, simultaneous neural machine translation (NMT) (Gu et al., 2016) that translates the source text into target text concurrently, and simultaneous text to speech (TTS) that synthesizes target speech given target text in real time. However, the cascaded approach suffers from two problems: 1) Simultaneous ASR, NMT and TTS models may perform well when trained separately, but cannot ensure good performance when cascaded together during inference, since the transcribed text by simultaneous ASR could be erroneous and will affect the accuracy of downstream models (Zhang et al., 2019); 2) Three-stage cascaded models cause more translation delay (the total delay is the sum of the delay of ASR, NMT and TTS), which is critical for simultaneous speech to speech translation.

Inspired by the recently proposed end-to-end full-sentence speech to speech translation (Jia et al., 2019), in this paper, we propose end-to-end simultaneous speech to speech translation called SimulS2S, which directly translates speech from source language into target language concurrently. SimulS2S adopts an encoder-decoder framework, which leverages the wait-k strategy (Ma et al., 2018) to schedule the listen (hear more source speech) and interpret (translate into target speech) decisions on the source and target speech respectively. Considering that the wait-k strategy requires both source and target speech sequence to be discrete segments, we further design a speech segmenter for speech segmentation. As a proof of concept, SimulS2S is supposed to alleviate the error propagation problem and reduce the delay of simultaneous translation, compared with the cascaded pipeline that trains simultaneous ASR, NMT and TTS models separately.

However, training an end-to-end simultaneous speech translation system is challenging, considering that simultaneous ASR, NMT and TTS are challenging tasks by themselves. Therefore, we introduce several techniques to assist the training of SimulS2S: 1) We utilize a curriculum learning mechanism to train the model gradually from full-sentence translation to simultaneous translation, with decreasing k in the wait-k strategy. 2) We introduce two auxiliary tasks including simultaneous ASR, S2T (speech to text translation) which share the same encoder with SimulS2S model to help the training of the encoder; 3) We leverage knowledge distillation to transfer the knowledge from the cascaded NMT and TTS models to SimulS2S model.

Experiments on Fisher Spanish-English conversation translation datasets demonstrate that SimulS2S: 1) achieves low translation delay and reasonable translation quality compared with full-sentence end-to-end speech to speech translation (without simultaneous translation), and 2) slightly underperforms the simultaneous translation with three-stage models in terms of both translation accuracy and delay. Considering the difficulty of the end-to-end simultaneous translation itself, there is big potential for future research.

The contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to propose an end-to-end simultaneous speech to speech translation system, and design a way to train it efficiently.
- We devise a curriculum learning mechanism to train the model gradually from full-sentence translation to simultaneous translation, by adjusting different k in wait-k strategy.
- We introduce two auxiliary tasks including simultaneous ASR and S2T, as well as a novel knowledge distillation from cascaded NMT and TTS models to help the training of SimulS2S model.
- Experiments on Fisher Spanish-English translation datasets demonstrate the effectiveness of SimulS2S as a proof of concept.

2 SimulS2S

In this section, we first describe the design of the proposed SimulS2S framework, and then introduce the key model of SimulS2S. We further introduce the curriculum learning mechanism and describe the auxiliary tasks and knowledge distillation to help the training of SimulS2S model.

The overview of SimulS2S framework is shown in Figure 1. The main task of SimulS2S leverages a source speech encoder and a target speech decoder both with a speech segmenter and a wait-k strategy, which are not shown in the figure and will be described in next subsection. The auxiliary simultaneous ASR and S2T task share the same encoder with SimulS2S model. The distillation task leverages a cascaded NMT and TTS (both are not simultaneous) as the teacher model, which is used to transfer the knowledge to SimulS2S model.

In the following, we describe the key model of SimulS2S in Section 2.1, and the curriculum learning mechanism in Section 2.2, and the auxiliary tasks and distillation task in Section 2.3.

2.1 The Model of SimulS2S

In this section, we introduce the key model of SimulS2S, including the problem formulation for simultaneous speech translation, the speech segmenter and the details of the model structure.

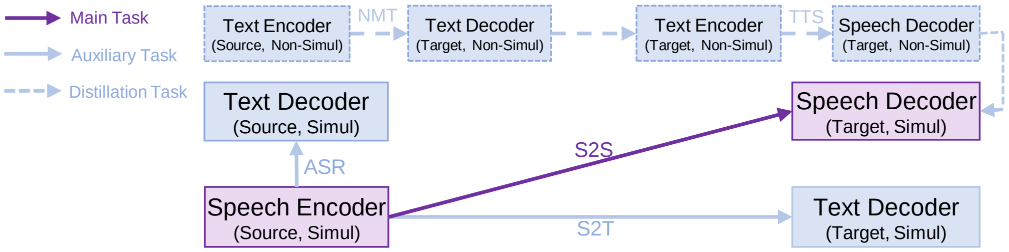


Figure 1: Overview of the SimulS2S framework. "Simul" and "Non-Simul" in the figure represent simultaneous translation and full-sentence translation respectively. The curriculum learning mechanism is not shown in the figure.

Problem Formulation Given a set of speech translation pairs $D = \{(x, y) \in (\mathcal{X} \times \mathcal{Y})\}$, the full-sentence translation model learns the parameter θ by minimizing the MSE loss $\mathcal{L}^{\text{full}} = \sum_{(x,y)\in D}(y - f(x;\theta))^2$, where $f(\cdot;\cdot)$ represents the prediction of mel-spectrogram if using mel-spectrogram as the acoustic feature. The loss $\mathcal{L}^{\text{full}}$ is usually calculated based on the autoregressive manner:

$$\mathcal{L}^{\text{full}} = \sum_{(x,y)\in D} \sum_{t=1}^{T_y} (y_t - f(y_{<t}, x; \theta))^2, \tag{1}$$

where $y_{<t}$ represents the frame of mel-spectrogram preceding position t and T_y is the number of frames in target speech y. In the full-sentence translation, the whole source speech sentence x can be seen for prediction. For simultaneous translation, we adapt the wait-k strategy (Ma et al., 2018) that is originally designed for simultaneous text to text translation into speech to speech translation. The loss function $\mathcal{L}^{\text{simul}}$ is

$$\mathcal{L}^{\text{simul}} = \sum_{(x,y)\in D} \sum_{t=1}^{T_y} (y_t - f(y_{<t}, x_{<\delta(t,k)}; \theta))^2, \tag{2}$$

where k corresponds to the wait-k strategy, $x_{<\delta(t,k)}$ represents the source speech preceding position $\delta(t,k)$. We describe how to design $\delta(t,k)$ in speech sequence. Instead of modeling the granularity of k in the frame level that does not contain enough semantic meaning, we model it in the segment level where a segment of speech represents a word or phrase, which is more suitable for translation. Denote $s(t)$ as the index of the speech segment which the t-th frame belongs to. In this way, $\delta(t,k)$ represents the start position of the $(s(t)+k)$-th source segment.

According to the formualtion in Equation 2, the model will wait for the first k source speech segments and then start to translate a target segment. After that, once receiving a new source segment, the decoder generates a new target segment until there is no more source segment, and then the translation degrades to the full-sentence translation.

Speech Segmenter However, Equation 2 assumes both the source and target speech sequence can be split into discrete segments, which is a non-trivial problem for speech sequence. In this paper, we propose a simple yet efficient method to split the speech mel-spectrogram into segments, where each segment is regarded as discrete tokens and represents a word or short phrase. Our method consists of the following steps: 1) We first extract the F0 and voice intensity from the audio sequence using Parselmouth (Jadoul et al., 2018). 2) We label a frame as unvoiced if its corresponding F0 can not be extracted and the voice intensity is below a threshold. 3) If the number of successive unvoiced frames exceeds a certain threshold, the unvoiced frames are used to split the audio into segments.

Model Structure The encoder of SimulS2S model follows the basic structure of Transformer based text to speech model (Li et al., 2019; Ren et al., 2019) but differs from the following modules: 1) In order to support the speech input in the encoder side, we use a pre-net which consists of multiple convolutional layers to extract features from mel-spectrograms; 2) We use a self-attention mechanism in the encoder with wait-k strategy; 3) Similarly, we use a decoder-to-encoder attention mechanism in the decoder with wait-k strategy. The attention masks of the two attention mechanism with wait-k strategy are designed according to Equation 2, which ensures that the source frames can only attend to the frames available at that time in the self-attention of the encoder, while the target frames can only attend to the source frames following the wait-k strategy.

2.2 Curriculum Learning for SimulS2S

In speech to speech translation, simultaneous translation is harder than full-sentence translation, while simultaneous translation with smaller k in the wait-k strategy is further harder than that with bigger k. Therefore, we design a curriculum learning mechanism (Bengio et al., 2009) by decreasing k in wait-k strategy (full-sentence translation can be regarded as $k = +\infty$), which gradually increases the difficulty of the training task.

Formally, the curriculum learning mechanism optimizes the simultaneous translation loss $\mathcal{L}^{\text{simul}}$ (defined in Equation 2) by gradually switching k from $+\infty$ to a desired K. We design several curriculum mechanisms to switch k, as shown in Figure 2. We put the detail descriptions of these mechanisms in Appendix.

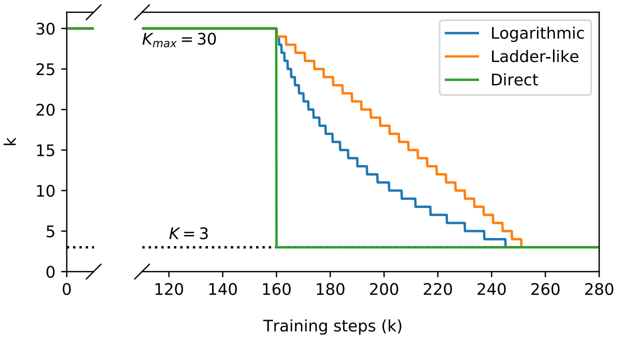


Figure 2: Different curriculum mechanisms to switch k from K_{max} to a desired K (e.g., 3). Ideally K_{max} should be $+\infty$. We just consider $K_{\text{max}} = 30$ in this work, which can cover the length of most speech sequence in terms of segments.

2.3 Auxiliary Tasks and Knowledge Distillation

Considering the difficulty of simultaneous speech to speech translation, in this section, we propose several techniques to boost the accuracy of SimulS2S, including auxiliary simultaneous ASR and S2T tasks, data-level knowledge distillation from cascaded NMT and TTS models.

Auxiliary Simultaneous ASR and TTS Multitask training is critical to improve the model performance of the speech to speech translation model (Jia et al., 2019). We add the auxiliary ASR and S2T tasks which leverage a simialr wait-k strategy and share the same encoder with the SimulS2S model, as shown in Figure 1. The decoders for both ASR and S2T follows the decoder in Vaswani et al. (2017b). ASR and S2T can help the shared speech encoder to extract more meaningful representations that is closer to the representation of source text, which will help speech to speech translation.

Knowledge Distillation from Cascaded NMT and TTS Model To further boost the accuracy of SimulS2S, we leverage data-level knowledge distillation to transfer the knowledge from a TTS model cascaded on an NMT model to the SimulS2S model, where both the NMT and TTS models are full-sentence but not simultaneous, as shown in Figure 1. Following Kim & Rush (2016); Tan et al. (2019), we use NMT teacher model to generate target text given source text that is paired with source speech, and further uses TTS model to synthesize targeted speech given translated target text, and then pair the synthesized target speech with source speech to train the SimulS2S model.

2.4 The Metric for SimulS2S

In this section, we describe the metrics to measure the translation delay and accuracy for simultaneous speech to speech translation.

Translation Delay Translation delay is an important metric for simultaneous translation. We use average proportion (AP) (Cho & Esipova, 2016) and average latency (AL) (Ma et al., 2018) to measure the delay. AP measures the averaged absolute delay incurred by each target token and AL measures the degree that the user is out of sync with the speaker. For simultaneous speech to speech translation task, we extend the AP and AL metric that is originally calculated on word sequence to speech sequence. Our extended AP is defined as

$$AP(x, y) = \frac{1}{|x||y|} \sum_{t=1}^{|y|} g(t), \tag{3}$$

where x and y are the source and target speech, $|x|$ and $|y|$ are the total time duration (in frames) of source and target speech, $g(t)$ is real-time delay (in frames) in terms of source speech when generating the t-th target speech frame. Our extended AL is defined as:

$$AL(x, y) = \frac{1}{\tau(|x|)|y|} \sum_{t=1}^{\tau(|x|)} max(g(t) - \frac{t-1}{r}, 0), \tag{4}$$

where $\tau(|x|)$ denotes the earliest timestep (in frames) where our model has consumed the entire source speech sequence:

$$\tau(|x|) = \arg\min_{t}(g(t) = |x|), \tag{5}$$

and $r = |y|/|x|$ is the duration ratio between target and source speech sequence.

Translation Accuracy To evaluate the accuracy of simultaneous speech to speech translation, we use a pre-trained ASR model to transcribe the generated target speech into text, and calculate the BLEU score (Papineni et al., 2002b) according to the ground-truth reference text translation following Jia et al. (2019). Due to potential recognition errors by ASR, this can be thought of as a lower bound of the underlying translation quality.

3 Experiments and Results

We first describe experimental settings, and then report the experiment results, and conduct some analyses on SimulS2S.

3.1 Experimental Settings

Datasets We conduct experiments on the Fisher Spanish-English corpus, which consists of Spanish telephone conversations and the corresponding English translations (Post et al., 2013). Fisher Spanish-English corpus contains 130k audio clips in source language and the corresponding transcripts in source and target language. We synthesize target speech from the target transcript using an English TTS system with a single (female) speaker. We use the official splits for train/dev/test set. For the speech data, we convert the raw waveform into mel-spectrograms following Shen et al. (2018) with 50 ms frame size and 12.5 ms frame hop. For the text data, we convert the transcriptions into characters.

Model Configuration We use Transformer (Vaswani et al., 2017b) as the basic SimulS2S model structure. We set the model hidden size, feed-forward hidden size, number of encoder and decoder-layers to 512, 2048, 6 and 6 respectively. Different from the 2-layer dense network in Transformer, we use a 2-layer 1D convolutional network (Gehring et al., 2017) with ReLU activation and left padding in the speech side (Ren et al., 2019). The motivation is that the adjacent hidden states are more closely related in the character and mel-spectrogram sequence in speech tasks. The filter size and kernel size of 1D convolution are 2048 and 9. The pre-net is a 3-layer left-padding convolutional network for both the encoder and decoder, and the output dimension equals to the hidden size of the encoder. In the speech segmenter, the thresholds of voice intensity and the number of successive unvoiced frames are 15db and 12.

Training and Inference We train the models with 4 NVIDIA Tesla V100 GPUs, each with batch size of roughly 8 sentences. We follow the default parameters of Adam optimizer (Kingma & Ba, 2014) and learning rate schedule in Vaswani et al. (2017a). We evaluate the translation quality by tokenized case sensitive BLEU (Papineni et al., 2002a) with multi-bleu.pl[1] after recognizing the translated speech in text using the Transformer based ASR model (Mohamed et al., 2019), which is trained on the 960 hours LibriSpeech corpus and Fisher corpus. We will release the code in Github when the paper is open to the public.

3.2 Experiment Results

Translation Accuracy We first evaluate the accuracy of SimulS2S model with different k in wait-k strategy. Besides regular evaluation with the same k both in training and inference, we also report the results of the model trained with k and test with different k' to observe the accuracy changes. The results are shown in Table 1. It can be seen that compared with the full-sentence translation (both k and k' is ∞), simultaneous translation does not drop much accuracy. We can also observe from this table that for a certain k' in inference, a smaller k in training usually results in better accuracy, mainly due to two reasons: 1) Training on more difficult task (small k) will make the model robust during inference; 2) Our curriculum mechanism ensures the model has already been trained on bigger k, which is consistent with bigger k during inference.

Train Test	k=1	k=3	k=5	k=7	k=9	k=∞
k'=1	**17.33**	12.04	9.09	6.82	6.11	4.68
k'=3	25.05	**25.13**	23.70	21.60	18.85	19.33
k'=5	27.46	**27.59**	27.52	26.26	24.66	23.83
k'=7	27.58	**29.19**	28.09	26.73	26.11	23.60
k'=9	26.86	28.62	28.01	**28.79**	27.88	24.69
k'=∞	24.56	25.04	25.17	24.85	25.30	**30.31**

Table 1: The BLEU scores of SimulS2S on the test set. The model is trained with wait-k but test with another wait-k'. The bold numbers are the best score in a row.

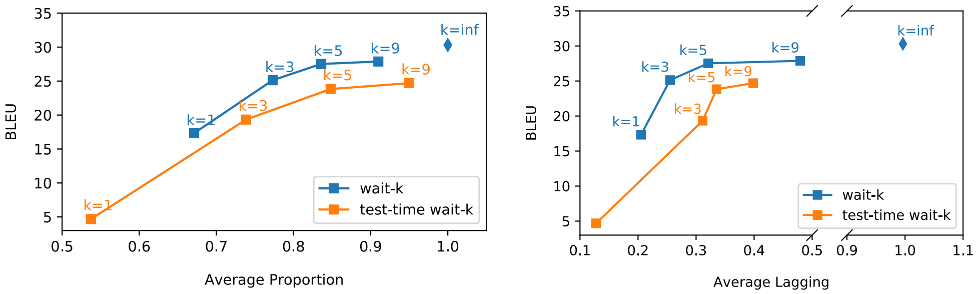
Translation Delay We plot the translation quality vs. translation delay of the wait-k model and test-time wait-k model in Figure 3a and 3b. The translation quality is measured by BLEU score while translation delay is measured by AP and AL, as denoted in Section 2.4. The wait-k model represents our SimulS2S with the same k both in training and inference, while the test-time wait-k model represents the model is trained with full-sentence translation but only test with wait-k strategy, which can be regarded as a naive baseline for simultaneous translation. We can see that the translation accuracy increases as k increases, with the sacrifice of translation delay. The accuracy of wait-k model is always better than the test-time wait-k, which demonstrates the effectiveness of the wait-k strategy during training.

Comparison with Cascaded Models We further compare SimulS2S with the three-stage method by cascading simultaneous ASR, NMT and TTS model together. We show the BLEU scores of the two methods under different k (both methods use the same k during training and inference) in Table 2. It can be seen that while SimulS2S performs slightly worse than cascaded method, the gap is close. Considering the difficulty of the end-to-end simultaneous translation itself, there is big potential for future research.

Model	k=1	k=3	k=5
Cascaded (ASR+NMT+TTS)	20.42	28.01	30.82
SimulS2S	17.33	25.13	27.52

Table 2: BLEU score comparison between three-stage cascaded method and SimulS2S.

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



(a) The translation quality against the latency in terms of AP.

(b) The translation quality against the latency in terms of AL.

Figure 3: The translation quality against the latency metrics (AP and AL).

3.3 Method Analyses

Curriculum Learning Strategy We conduct experiments to evaluate the effectiveness of different curriculum learning strategies under different k in wait-k strategy. The results are shown in Table 3. It can be seen that the models with curriculum learning strategies (Direct, Ladder-like and Logarithmic) achieve better accuracy than that without curriculum learning, which demonstrates the effectiveness of our proposed curriculum learning mechanism. Among the different curriculum strategies, the logarithmic strategy performs the best in most cases. Therefore, we use Logarithmic as the default curriculum strategy in the experimental study of this work, unless otherwise stated.

Model	k=1	k=3	k=5	k=7	k=9
Without CL	3.46	6.18	11.01	14.65	14.45
Direct	15.24	23.89	26.06	**27.51**	26.96
Ladder-like	15.60	23.25	26.46	27.40	27.12
Logarithmic	**17.33**	**25.13**	**27.52**	26.73	**27.88**

Table 3: Comparison of different curriculum learning strategies.

Auxiliary Tasks We evaluate the effectiveness of the ASR and S2T auxiliary tasks. As shown in Table 4, starting from the naive simultaneous speech to text translation model without any auxiliary task (denoted as S2S), we add the auxiliary tasks of simultaneous ASR (Row 3) and S2T (Row 4). Without any auxiliary tasks, the naive S2S model cannot achieve reasonable accuracy, which demonstrates the difficulty of simultaneous speech to speech translation. The translation accuracy can be boosted by adding ASR auxiliary task, and the gains become larger if further adding S2T at the same time.

Model	k=1	k=5	k=9
S2S	0.06	0.08	0.09
+ASR	3.87	11.23	12.21
+ASR+S2T	17.33	27.52	27.88

Table 4: Ablation study on auxiliary tasks.

Speech Segmentation We evaluate the precision of our segmentation method using average absolute error (ASE) (Mesaros & Virtanen, 2008) on the test set, which measures the alignment errors between the segmentation of our method and the ground-truth in terms of time (ms). We use forced alignment tool[2] to obtain the ground-truth word boundaries on the test set. The predicted word

[2] https://github.com/lowerquality/gentle

boundary by our segmentation method is matched to the nearest ground-truth boundary to calculate the ASE. Considering the number of predicted word boundaries by our method is usually fewer than the ground-truth boundaries, due to that our method prefers not to segment two words that with no silence frame in between, we further count the metric of missing rate, which denotes the ratio of the number of boundaries that are not detected by our method. Our segmentation method achieve 110ms ASE and 23.4% missing rate, which demonstrates the effectiveness of our segmentation method. A high missing rate means a speech segment will cover more words, which influences the delay of simultaneous translation. However, we can adjust k in wait-k strategy to compensate the delay.

4 Related Work

In this section, we introduce the related works for simultaneous speech to speech translation, including full-sentence speech translation, simultaneous translation and curriculum learning.

Full-sentence Speech Translation Speech translation has been a hot research topic in the field of artificial intelligence (Lavie et al., 1997; Nakamura et al., 2006; Hori et al., 2009; Wahlster, 2013; Weiss et al., 2017; Bansal et al., 2018; Sperber et al., 2019; Zhang et al., 2019; Jia et al., 2019). Early works on speech to speech translation rely on a three-stage method by cascading ASR, NMT and TTS models (Lavie et al., 1997; Nakamura et al., 2006; Hori et al., 2009; Wahlster, 2013) or focus on end-to-end speech to text translation (Weiss et al., 2017; Bansal et al., 2018; Liu et al., 2019; Zhang et al., 2019). Recently, Jia et al. (2019) propose a fully end-to-end speech translation model that translate from source speech into target speech directly. However, previous works on speech to speech translation, with either cascaded models or end-to-end models, focus on full-sentence translation, which introduced too much delay when used in spoken lectures or conversations. In this work, we study end-to-end speech to speech translation in the simultaneous scenario to reduce the delay in translation without loss much of translation accuracy.

Simultaneous Translation Simultaneous translation (Bérard et al., 2016; Weiss et al., 2017; Liu et al., 2019) is widely considered as one of the challenging tasks for spoken language translation Fügen et al. (2007); Oda et al. (2014); Dalvi et al. (2018). Previous works on simultaneous speech to speech translation leverage a three-stage cascaded models: simultaneous ASR, NMT and TTS, and most research works focus on simultaneous NMT (Gu et al., 2016; Ma et al., 2018; Zheng et al., 2019). Most works on simultaneous NMT leverage a schedule to balance the translation delay against translation quality by deciding whether to read source tokens (see more source words) or write target tokens (translate into target words). Ma et al. (2018) introduced a very simple but effective wait-k strategy. In this paper, we also leverage wait-k strategy for simultaneous translation, but with adaptation into speech sequence. Different from the previous works that leverage three-stage cascaded models, we tackle it in an end-to-end way, which has the potential to eliminate error propagation and reduce translation delay in cascaded models.

Curriculum Learning Humans usually learn better when the curriculums are organized from easy to hard. Inspired by that, Bengio et al. (2009) proposed curriculum learning, a machine learning training strategy that feeds training instances to the model from easy to hard. Most of the works on curriculum learning focus on the determining the orders of data (Lee & Grauman, 2011; Sachan & Xing, 2016) or task (Pentina et al., 2015; Sarafianos et al., 2017). In our setting, we design curriculums for neither data samples nor tasks, but the training mechanisms by adjusting different k in the wait-k strategy.

5 Conclusion

In this work, we have proposed SimulS2S, an end-to-end simultaneous speech to speech translation system, which consists of a speech encoder, a speech decoder both with a speech segmenter and wait-k strategy. We further introduce a curriculum mechanism, two auxiliary tasks and knowledge distillation to boost the accuracy of SimulS2S. Experiments on Fisher Spanish to English corpus demonstrate that SimulS2S achieves low translation delay and reasonable translation quality compared with the full-sentence end-to-end speech to speech translation (without simultaneous translation). As a proof of concept, our work shows that there is big potential for future research, considering the difficulty of the end-to-end simultaneous translation itself.

For future work, we will further improve the accuracy of simultaneous speech translation to approximate the accuracy of cascaded method and full-sentence translation.

REFERENCES

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 437–445. Association for Computational Linguistics, 2012.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Low-resource speech-to-text translation. *arXiv preprint arXiv:1803.09164*, 2018.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48. ACM, 2009.

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

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

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. Incremental decoding and training methods for simultaneous translation in neural machine translation. *arXiv preprint arXiv:1806.03661*, 2018.

Christian Fügen, Alex Waibel, and Muntsin Kolss. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4):209–252, 2007.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, pp. 1243–1252. JMLR. org, 2017.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. Learning to translate in real-time with neural machine translation. *arXiv preprint arXiv:1610.00388*, 2016.

Chiori Hori, Sakriani Sakti, Michael Paul, Noriyuki Kimura, Yutaka Ashikari, Ryosuke Isotani, Eiichiro Sumita, and Satoshi Nakamura. Network-based speech-to-speech translation. In *nnnn*, 2009.

Yannick Jadoul, Bill Thompson, and Bart De Boer. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15, 2018.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

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

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. Janus-iii: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 99–102. IEEE, 1997.

Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR 2011*, pp. 1721–1728. IEEE, 2011.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and M Zhou. Neural speech synthesis with transformer network. AAAI, 2019.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*, 2019.

Mingbo Ma, Liang Huang, Hao Xiong, Kaibo Liu, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, and Haifeng Wang. Stacl: Simultaneous translation with integrated anticipation and controllable latency. *arXiv preprint arXiv:1810.08398*, 2018.

Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

Takashi Mieno, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Speed or accuracy? a study in evaluation of simultaneous speech translation. In *Sixteenth Annual Conference of the ISCA*, 2015.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376, 2006.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Optimizing segmentation strategies for simultaneous speech translation. In *ACL*, pp. 551–556, 2014.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002a. URL `http://www.aclweb.org/anthology/P02-1040.pdf`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002b.

Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5492–5500, 2015.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus. In *Proc. IWSLT*, 2013.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *ASRU*, pp. 193–199. IEEE, 2017.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019.

Mrinmaya Sachan and Eric Xing. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 453–463, 2016.

Nikolaos Sarafianos, Theodore Giannakopoulos, Christophoros Nikou, and Ioannis A Kakadiaris. Curriculum learning for multi-task classification of visual attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2608–2615, 2017.

Anoop Sarkar. The challenge of simultaneous speech translation. In *PACLIC*, 2016.

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

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Attention-passing models for robust and data-efficient end-to-end speech translation. *TACL*, 7:313–325, 2019.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS 2017*, pp. 6000–6010, 2017a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017b.

Wolfgang Wahlster. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

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

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. Lattice transformer for speech translation. *arXiv preprint arXiv:1906.05551*, 2019.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. Simultaneous translation with flexible policy via restricted imitation learning. *arXiv preprint arXiv:1906.01135*, 2019.

A Appendix

A.1 Attention Mask in Simultaneous Translation

According to the wait-k strategy described above, we design the wait-k encoder self-attention and wait-k encoder-decoder attention mask as follows:

$$M^{\text{enc}}(i,j) = \begin{cases} 0, i < \ j \vee j < k \\ -\infty, \ \text{otherwise} \end{cases}, \tag{6}$$

$$M^{\text{enc-dec}}(i,j) = \begin{cases} 0, j \leq i + k \\ -\infty, \ \text{otherwise} \end{cases}, \tag{7}$$

When $M(i,j)$ equals to $-\infty$, the corresponding position in softmax output will approach zero, which prevents position i from attending to position j.

A.2 The Curriculum Mechanism for Training SimulS2S

Pacing Functions	Description
Ladder-like	$k_{\text{ladder}}(i) = \begin{cases} k_{max}, i < S \\ max(k_{max} - \frac{i-S}{3500}, K), i \geq S \end{cases}$
Logarithmic	$k_{\text{log}}(i) = \begin{cases} k_{max}, i < S \\ max(k_{max} - 8 * log_2 \frac{i-S}{10000} + 1, K), i \geq S \end{cases}$
Direct	$k_{\text{direct}}(i) = \begin{cases} k_{max}, i < S \\ K, \ \text{otherwise} \end{cases}$

Table 5: The proposed different curriculum mechanisms and their definitions.

A.3 Model Structure for SimulS2S

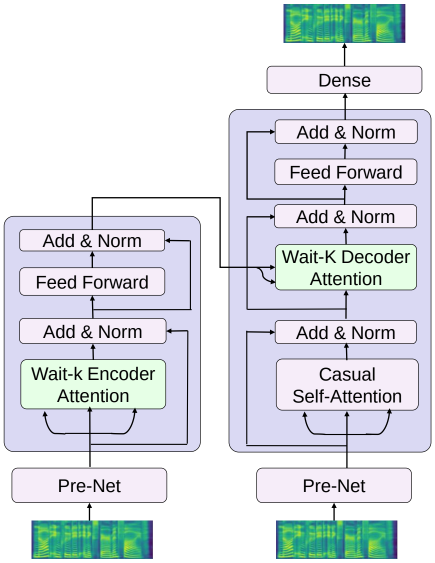


Figure 4: The encoder and decoder of the SimulS2T model.