

Simultaneous Speech-to-Speech Translation System with Transformer-based Incremental ASR, MT, and TTS

Ryo Fukuda, Sashi Novitasari, Yui Oka, Yasumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Tomoya Yanagita, Sakriani Sakti, Katsuhito Sudoh, Satoshi Nakamura

Nara Institute of Science and Technology (NAIST)

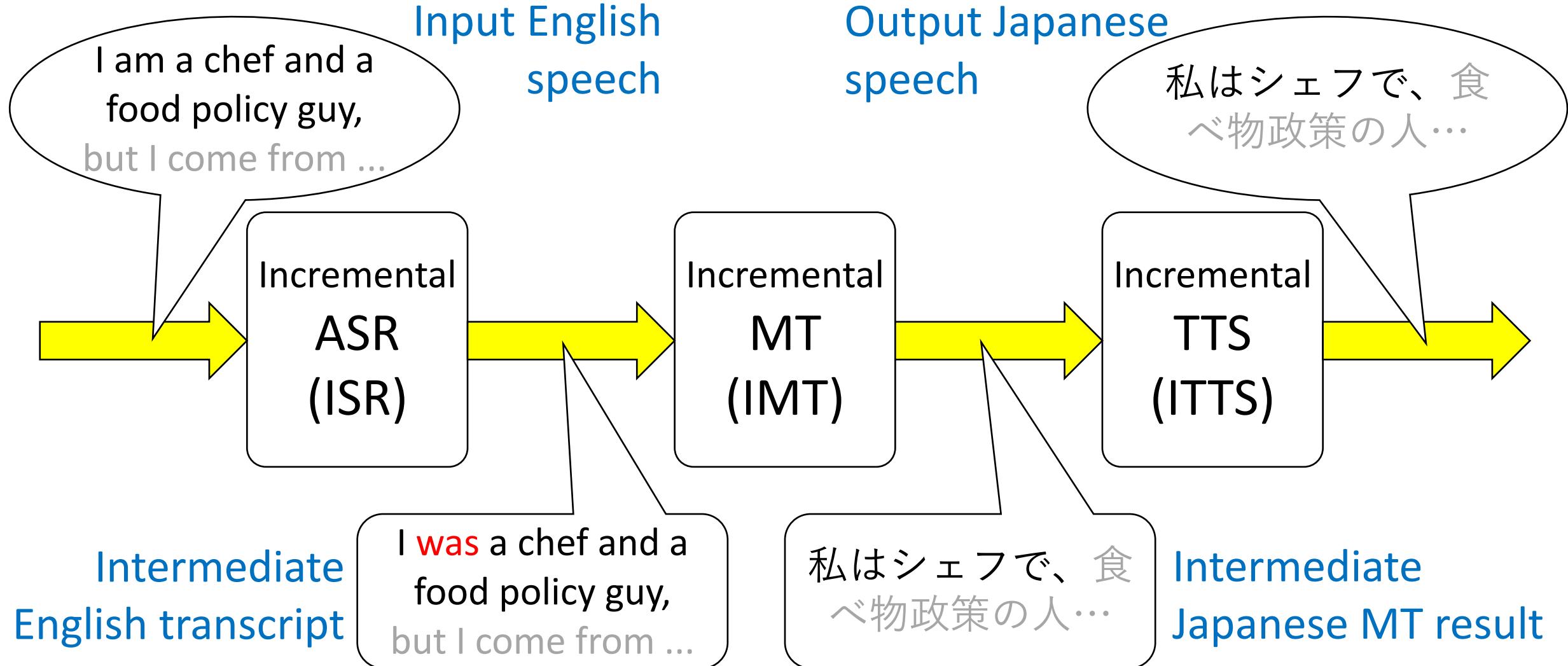
Video



Background

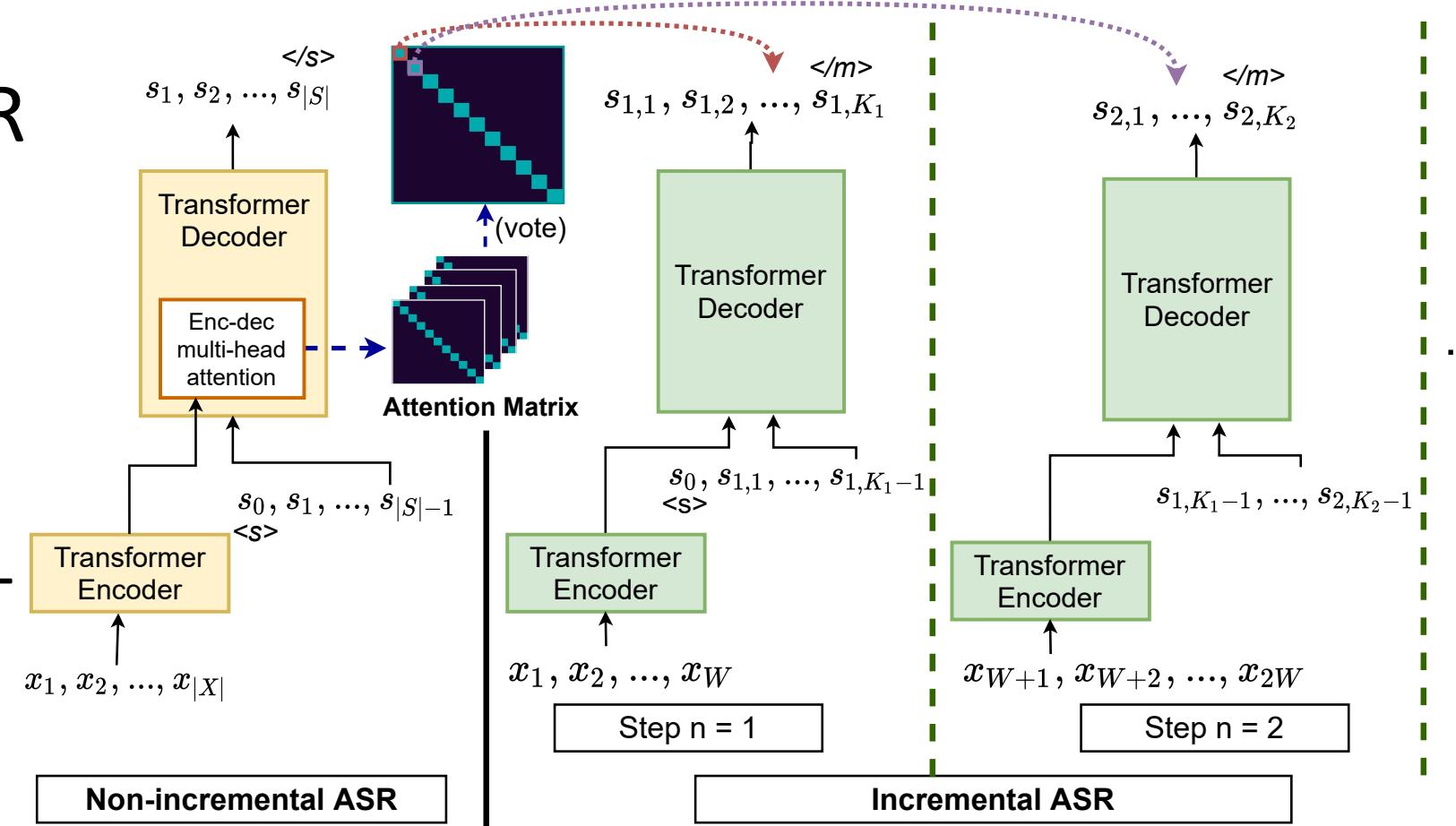
- Simultaneous Speech-to-Speech Translation from En to Ja
 - *Simultaneous* Translation
 - Not waiting for the end of a sentence; Real-time communication
 - *Speech-to-Speech* Translation
 - No displays for closed captions / subtitles
 - Simultaneous *Translation*
 - No interpretation efforts like summarization
 - *From English to Japanese*
 - Syntactically distant language pair; Large difference in word order

Cascade Simultaneous S2S Translation System



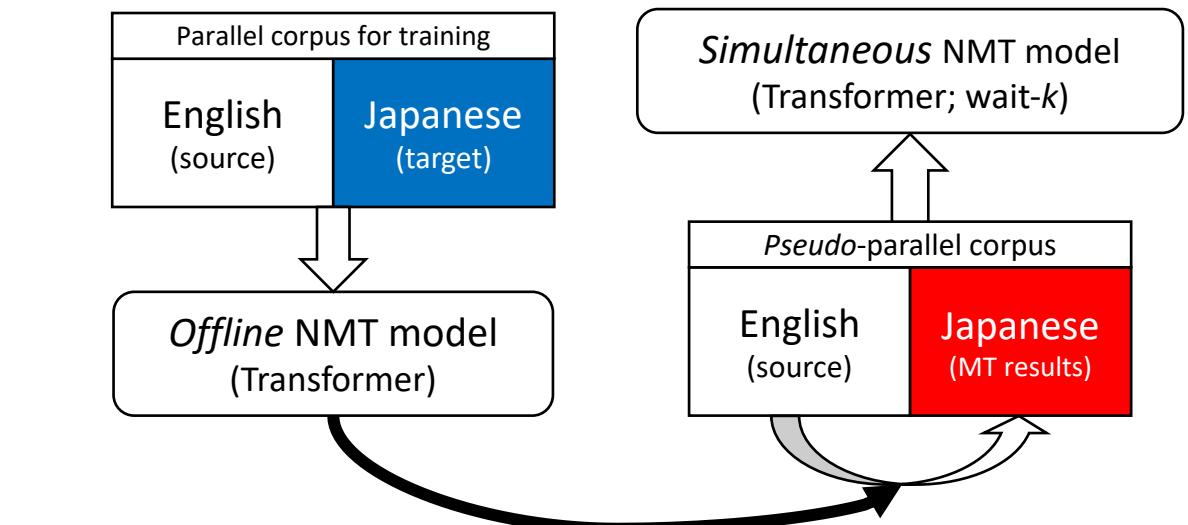
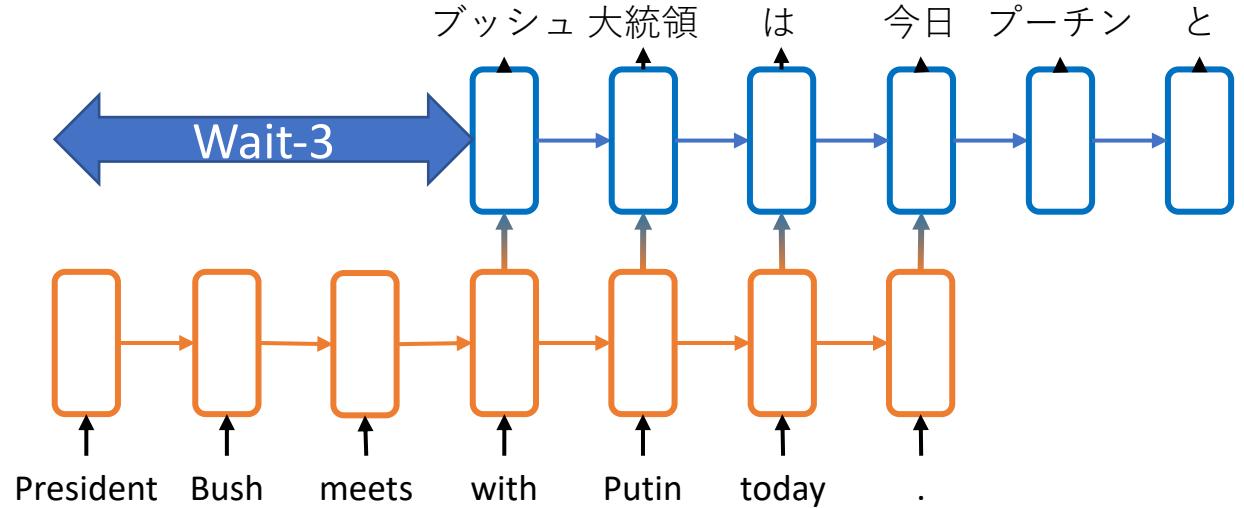
Incremental Speech Recognition (ISR)

- Transformer-based extension to our ISR using knowledge distillation (Novitasari+ 2019)
 - The student model mimics the speech-text alignment



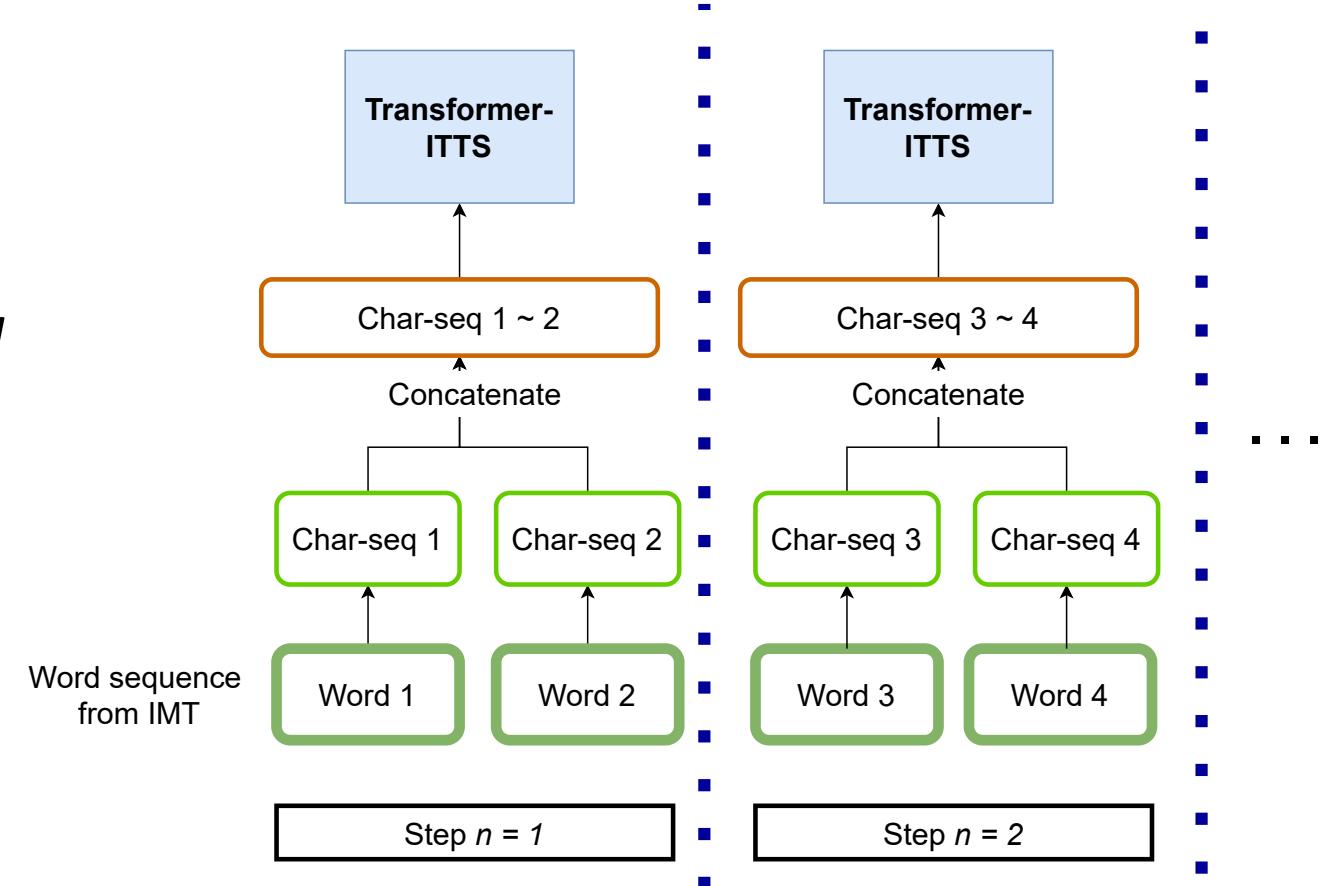
Incremental Machine Translation (IMT)

- Applies our text-to-text IMT (Fukuda et al. 2021)
 - Wait-k
 - Delays the decoding process in k input tokens
 - Sequence-level Knowledge Distillation
 - Trains the IMT model using offline MT results
 - Japanese chunk shuffling (details in the paper)



Incremental Text-to-Speech Synthesis (ITTS)

- Based on Transformer TTS (Li et al. 2019)
 - Takes fixed # of Ja words and convert them into *kana* (Japanese phonograms)
 - Predicts Mel-spectrogram
 - Generates speech signals using CBHG and Griffin-Lim
 - Continues until the end-of-sentence



Evaluation Overview

- The cascade system was evaluated using TED Talks data in:
 - System-level latency
 - Ear-Voice Span (EVS)
 - Span between the start of the input & output
 - Module-level quality
 - ISR: Word Error Rate (WER)
 - ISR+IMT: BLEU
 - ITTS: L2-norm loss and subjective evaluation (AB preference test)
 - Three latency regimes: low, medium, high (following IWSLT)

Evaluation: Latency

Latency regime	ISR delay (sec.)	ISR+IMT delay (sec.)	Total delay (=EVS) (sec.)	Hyper-parameters
Low	0.93	4.69	8.81	ISR: 64 frames IMT: k=10 ITTS: 5 words
Medium	0.93	8.43	11.87	ISR: 64 frames IMT: k=20 ITTS: 5 words
High	1.30	11.47	16.91	ISR: 64 frames IMT: k=30 ITTS: 7 words

Evaluation: Quality (ASR & MT)

- ISR

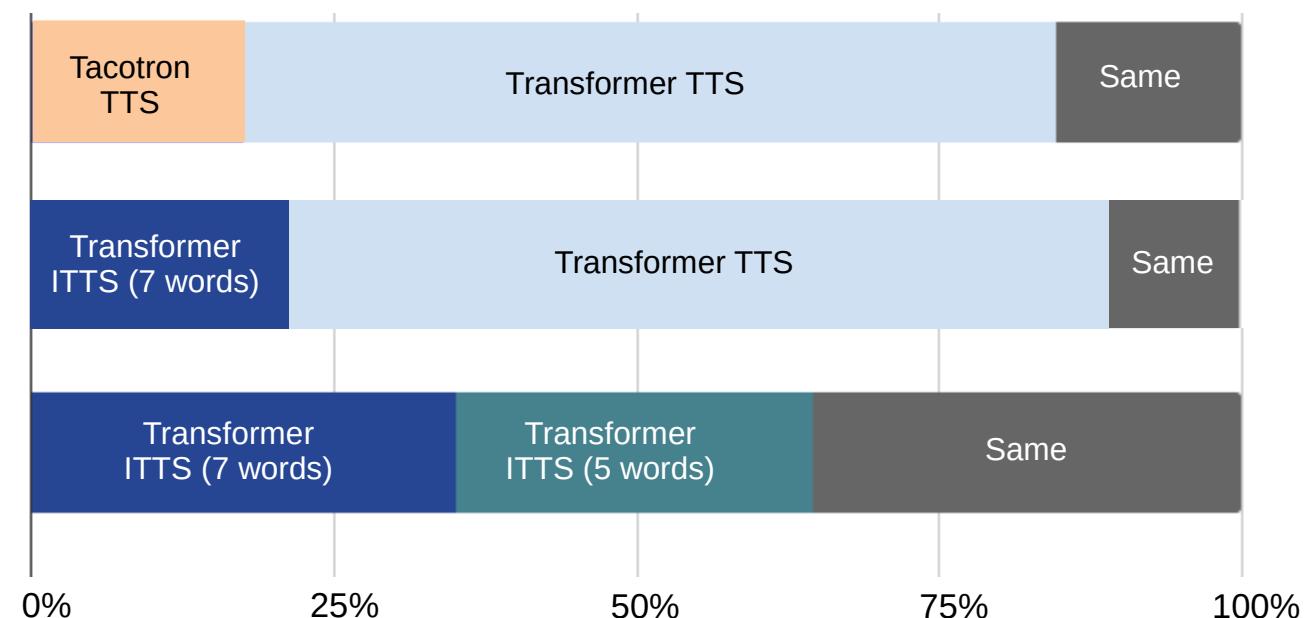
	Model	WER (%)
Non-incremental	LSTM	25.46
	Transformer	20.74
	LSTM (low latency)	31.88
	LSTM (high latency)	32.43
	Transformer (low latency)	32.06
	Transformer (high latency)	25.01

- ISR+IMT

	BLEU-4 (%)	Subjective evaluation	
		Adequacy	Fluency
Gold transcript+MT	15.7	3.41	3.93
Non-incremental ASR+MT	12.8	3.20	4.01
IMT (low latency)	5.1	2.80	3.03
IMT (medium latency)	8.4	2.98	3.54
IMT (high latency)	9.4	3.34	3.80

Evaluation: Quality (TTS)

	Model	L2-norm loss
Non-incremental	Tacotron	0.57
	Transformer	0.51
	Tacotron (low latency)	0.77
Incremental	Tacotron (high latency)	0.58
	Transformer (low latency)	0.65
	Transformer (high latency)	0.57



Conclusions

- Our En-to-Ja simultaneous speech-to-speech translation system using Transformer-based incremental modules
 - Module-level delays are still problematic
 - Suffers from speaking latency for long TTS output
 - Delay increases when a new output come during speaking
- Future work includes:
 - Aggressive anticipation to reduce the delay
 - Controlling the speaking duration
 - Evaluation of simultaneous speech-to-speech translation

References

- Novitasari, S. et al. 2019, Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition, Proc. Interspeech 2019
- Fukuda, R. et al., 2021, NAIST English-to-Japanese Simultaneous Translation System for IWSLT 2021 Simultaneous Text-to-text Task, Proc. IWSLT 2021
- Li, N. et al., 2019, Neural Speech Synthesis with Transformer Network, Proc. AAAI 2019