Inter-connection: Effective Connection between Pre-trained Encoder and Decoder for Speech Translation

Yuta Nishikawa[1], Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology, Japan

Major Findings

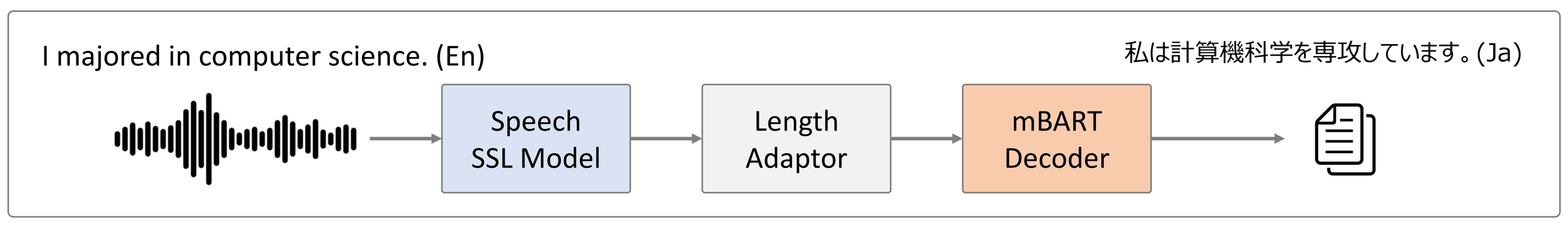
Inter-connection: weighted-sum aggregation improves speech translation

➢ Efficient in terms of parameter size

Background

A method combining self-supervised learning (SSL) models of speech with mBART decoder shows high performance in end-to-end speech translation (ST)

[Pham+2022] [Tsiamas+2022]

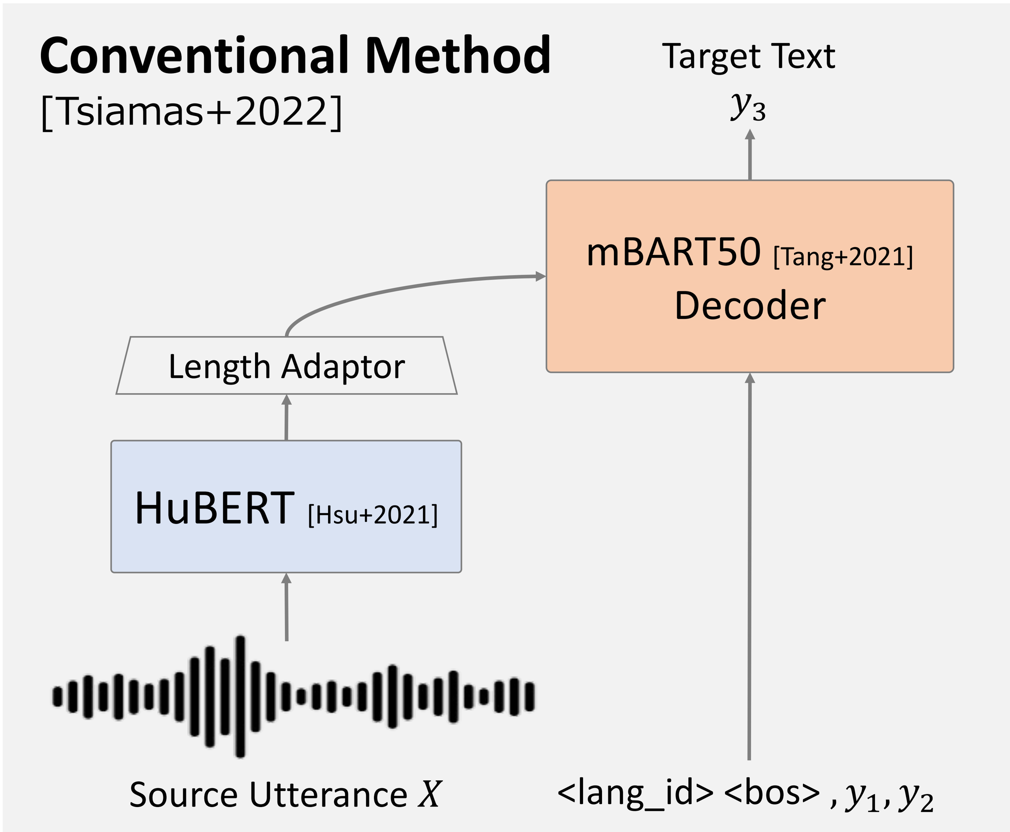


The simple connection method of the encoder-decoder model is **unable to utilize the information from speech SSL models.**

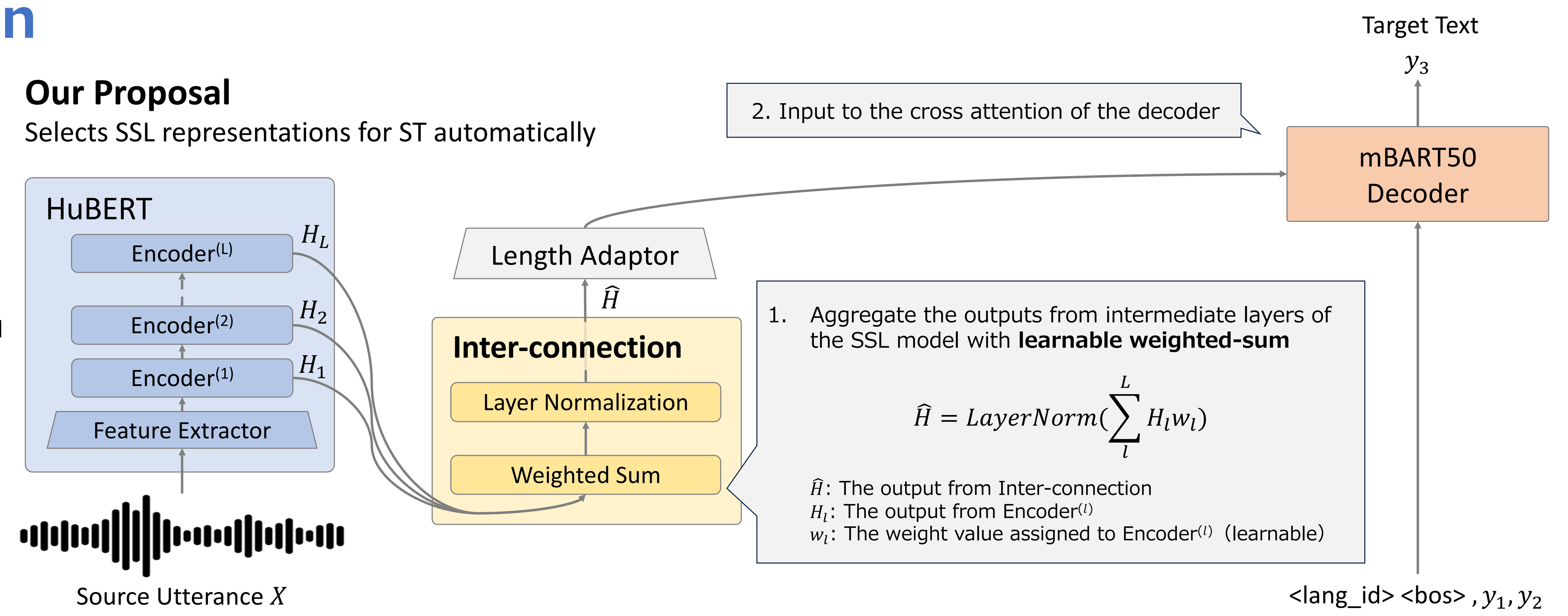
Features of speech SSL models [Pasad+2021]
- Autoencoder-like behavior
- Contains a lot of **useful information in intermediate layers** (phonetic or linguistic features)

Purpose ➡ Extracting and utilizing the SSL representations important for ST

Methodology: Inter-connection



Fully utilize the SSL model



Results and Analysis

Translation Quality

Motivation: Investigate whether Inter-connection improves ST

➢ Train the **multilingual model (En-De, Ja, Zh)** by MuST-C v2
➢ w/ and w/o parameter freezing of HuBERT

Evaluation Results on tst-COMMON

Model	BLEU↑			
	En-De	En-Ja	En-Zh	Ave.
w/ Parameter Freezing				
Baseline [Tsiamas+2022]	24.68	11.86	20.55	19.03
Inter-connection (Proposal)	**26.79**	**14.15**	**22.20**	**21.05**
w/o Parameter Freezing				
Baseline [Tsiamas+2022]	30.48	15.81	**24.82**	23.70
Inter-connection (Proposal)	**30.67**	**16.22**	24.59	**23.82**

2.02 BLEU↑

0.12 BLEU↑

Inter-connection improves ST in most language pairs

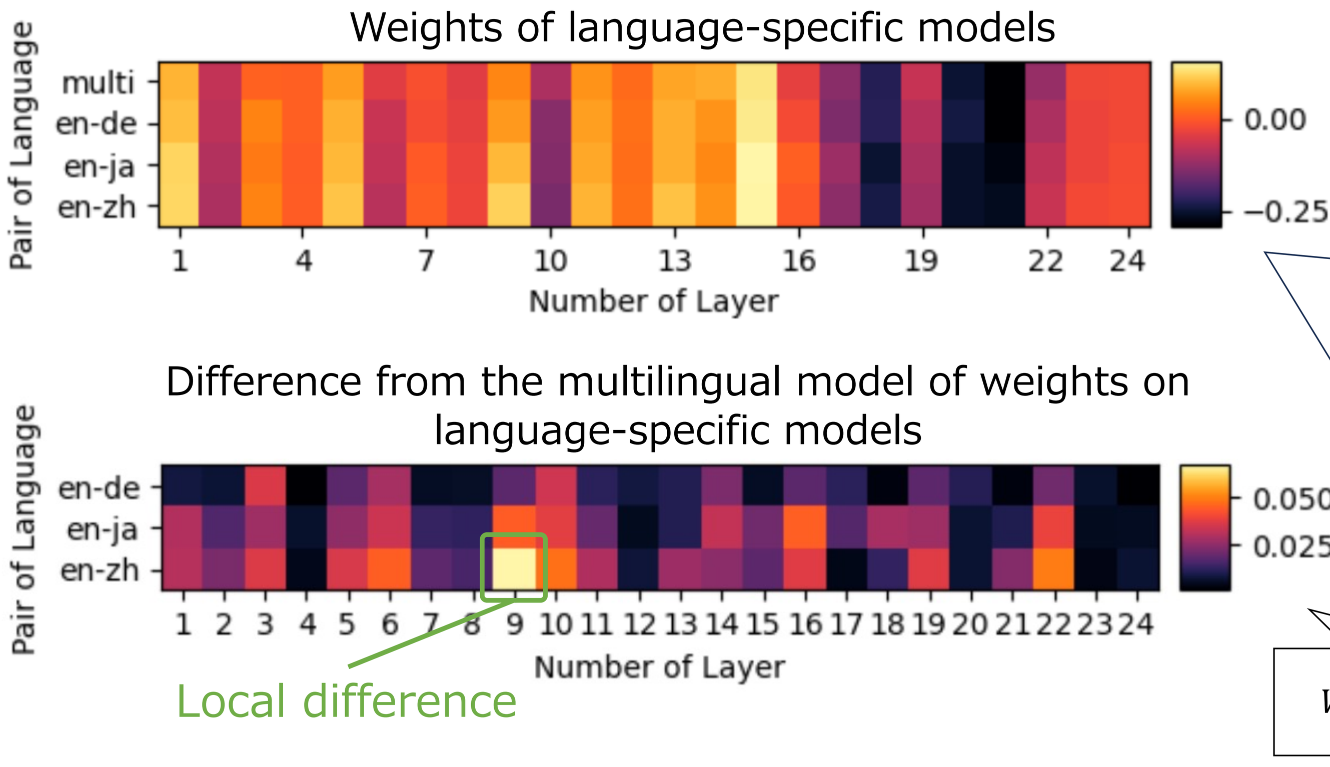
However, En-Zh w/o parameter freezing was not improved

Why?

Layer-wise Analysis

Motivation: Find out why performance dropped in En-Zh

➢ Train **bilingual models** for each language pairs (En-De, En-Ja, En-Zh)
➢ Compare weights of inter-connection between **bilingual model** and **multilingual model**



Cosine similarity with multilingual model

Lang-pair	Similarity
En-De	0.991
En-Ja	0.979
En-Zh	0.969

Most far

$$W_{diff} = |W_{multi} - W_{bi}|$$

➢ En-Zh weights are far from multilingual model
➢ Sharing weights across all language pairs might have a **negative effect**

Parameter-size Analysis

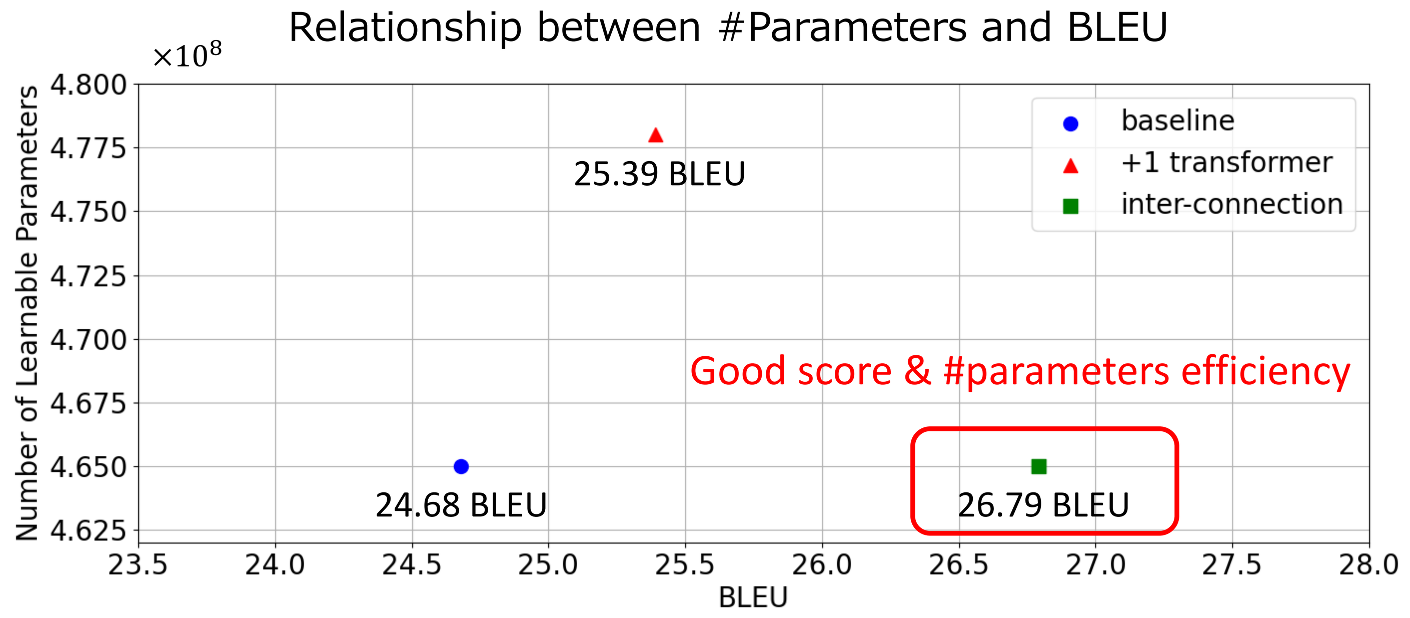
Motivation: Verify how efficient in terms of #parameters

#Parameters increased for each module

Module	#Parameters
+ 1 transformer	12M
Inter-connection	2K

Compare the performance in En-De w/ Parameter Freezing

Relationship between #Parameters and BLEU



Conclusion

Summary
- Aggregation by weighted-sum improves performance of ST
- In the multilingual model, sharing weights has a negative impact
- Efficient in terms of increasing number of parameters

Future works
- Reduce the negative effect of weight sharing
- Application and analysis for other tasks (e.g. ASR)

References

[Pham+2022] "Effective Combination of Pretrained Models – KIT@IWSLT2022", IWSLT2022
[Tsiamas+2022] "Pretrained Speech Encoders and Efficient Fine-tuning Methods for Speech Translation", IWSLT2022
[Hsu+2021] "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021
[Tang+2021] "Multilingual Translation from Denoising Pre-training", ACL-IJCNLP 2021
[Pasad+2021] "Layer-wise Analysis of a Self-Supervised Speech Representation Model", ASRU2021