音声翻訳研究の現状と今後

Satoshi Nakamura
With
Katsuhito Sudo and AHC ST-Team
Nara Institute of Science and Technology (NAIST), Japan
S-nakamura@is.naist.jp

Part of this work is supported by JSPS KAKENHI (Grant numbers JP21H05054, 21H03500, 17H06101)

My Speech Translation Research

1986 — ATR

1992 — NAIST

1999 — ATR

2006 — ATR

2008 — NICT

2010 — NICT

2011 — NAIST

2023

Read Speech

- Syntactically correct
- Clear utterance
- Limited domain
  Ex. “Conference Registration”

Daily Conversation

- Standard expression
- Unclear utterance
- Limited domain
  Ex. “Hotel Reservation”

Wider and Real Domain

- Wider and real domain
  “International Travel”
- Realistic expressions
- Noisy speech
- J-E, J-C speech translation

+ More languages for translation

- 21 multilateral text translation

A-STAR

- Multilateral translation for 8 Asian languages
- Network-based S2ST

C-STAR

- Multilateral translation for 7 world languages

IWSLT

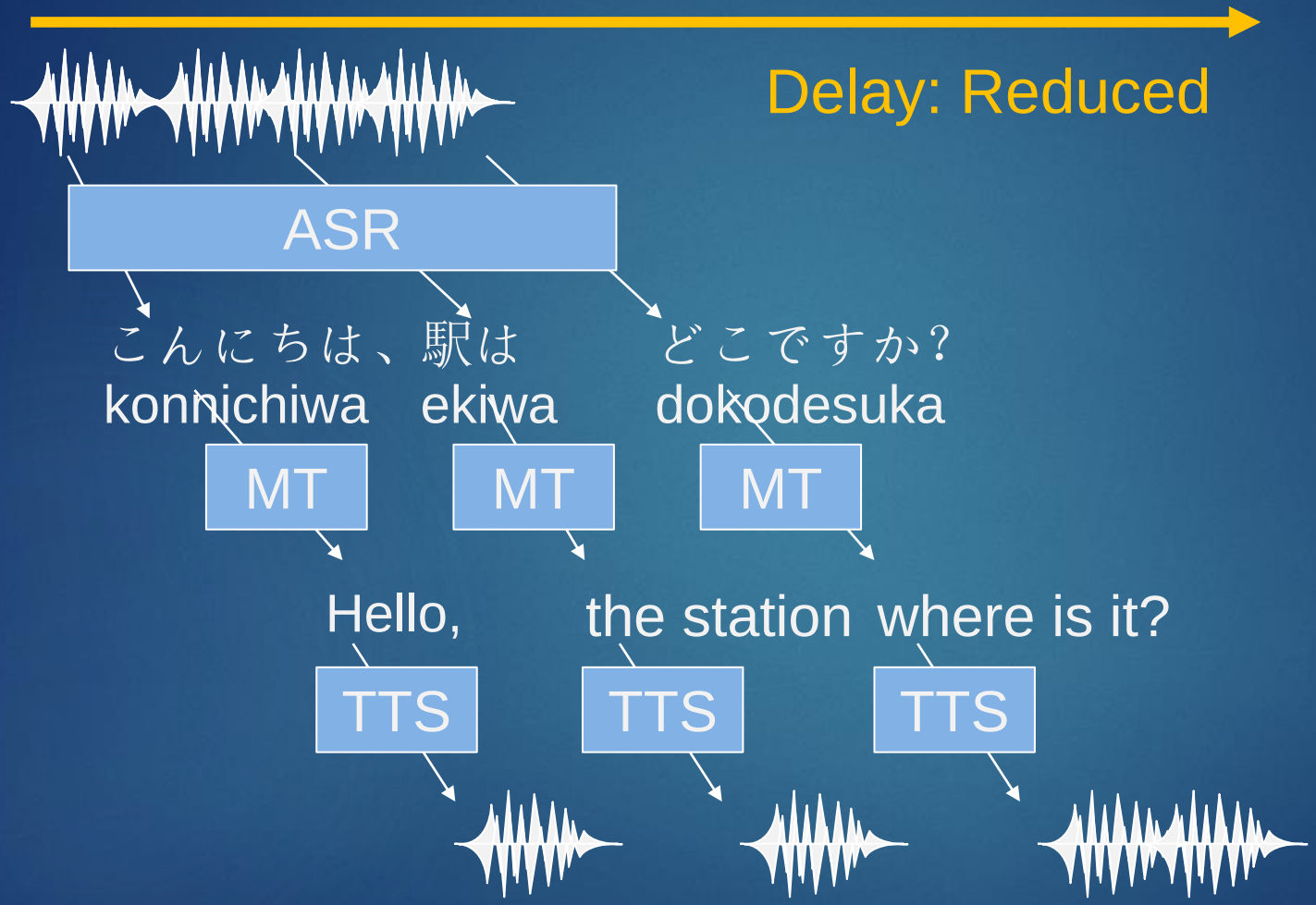
- Evaluation Campaign of S2S technologies

Simultaneous Speech-to-speech Translation @IWSLT

Approach

Rule-based Machine Translation → Statistical Machine Translation → Neural Machine Translation

Simultaneous Incremental Speech Translation



But, this is not easy!

Challenge in En-Ja Simul MT

- Long distance reordering

Source:

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplied (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

Example from Mizuno, A. "Simultaneous Interpreting and Cognitive Constraints," Journal of College of Literature, Aoyama Gakuin University, vol. 58, pp. 1–28, 2016

Typical Challenge in En-Ja Simul MT

Source:

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplied (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

Target:

(1) *kyuuen tan'tousha wa* (9) *ikiru tame no* (8) *shokuryo o motomete* (7) *mura o arashi mawatte iru* (6) *tairyo no nan-min tachi no* (5) *sewa o suru tame no* (4) *juubun na shokuryo ya mizu, shukuhaku shisetsu, iyakuhin ga* (3) nai to (2) ***itte imasu***.

Example from Mizuno, A. "Simultaneous Interpreting and Cognitive Constraints," Journal of College of Literature, Aoyama Gakuin University, vol. 58, pp. 1–28, 2016

Interpreters' *monotonic* translation

Source:

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplied (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

Target:

(1) *kyuuen tan'tousha tachi no* (2) ***hanashi de wa***, (4) *shokuryo, mizu, shukuhaku shisetsu, iyakuhin ga* (3) *tari zu* (6) *tairyo no nan-min tachi no* (5) sewa ga dekinai to no koto desu.

(7) *nan-min tachi wa ima muramura o arashi mawatte,* (9) *ikiru tame no* (8) *shokuryo o motomete iru no desu*.

Example from Mizuno, A. "Simultaneous Interpreting and Cognitive Constraints," Journal of College of Literature, Aoyama Gakuin University, vol. 58, pp. 1–28, 2016

Translation vs Interpretation

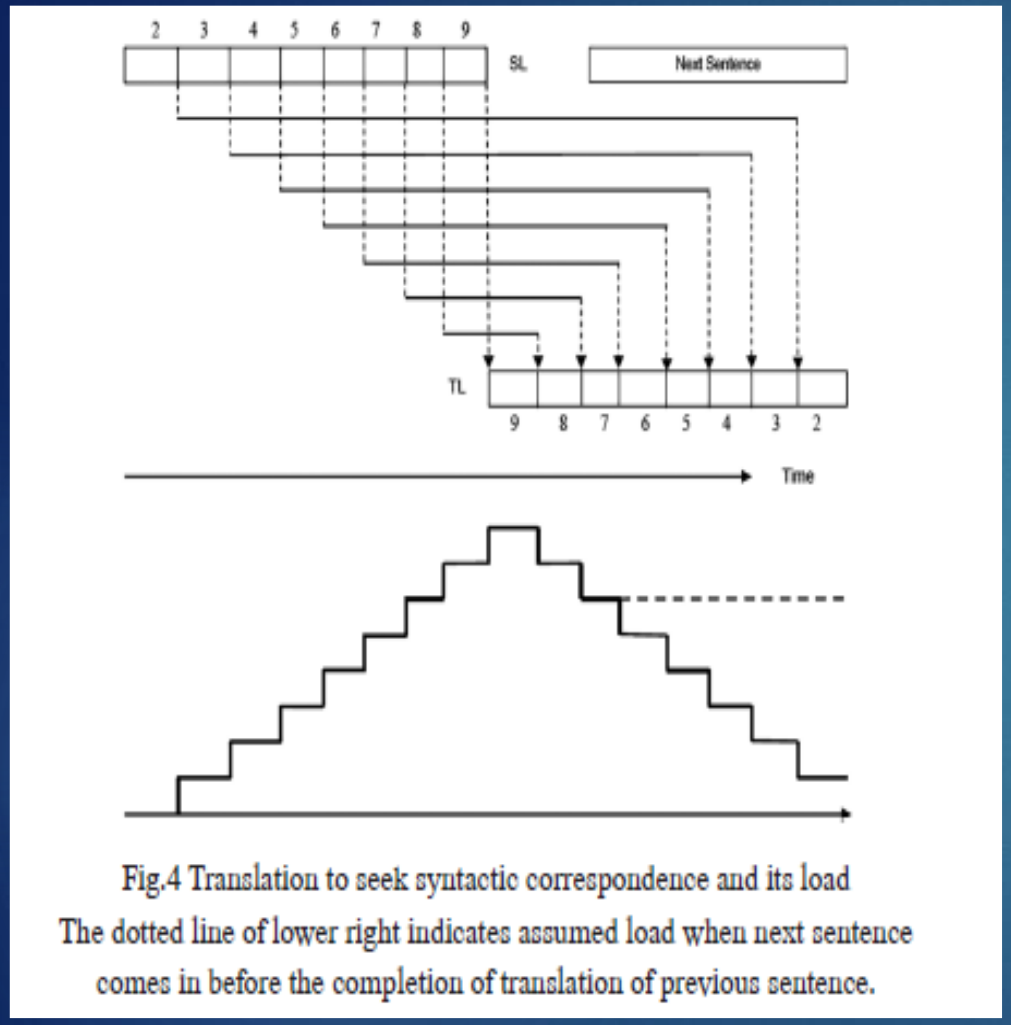
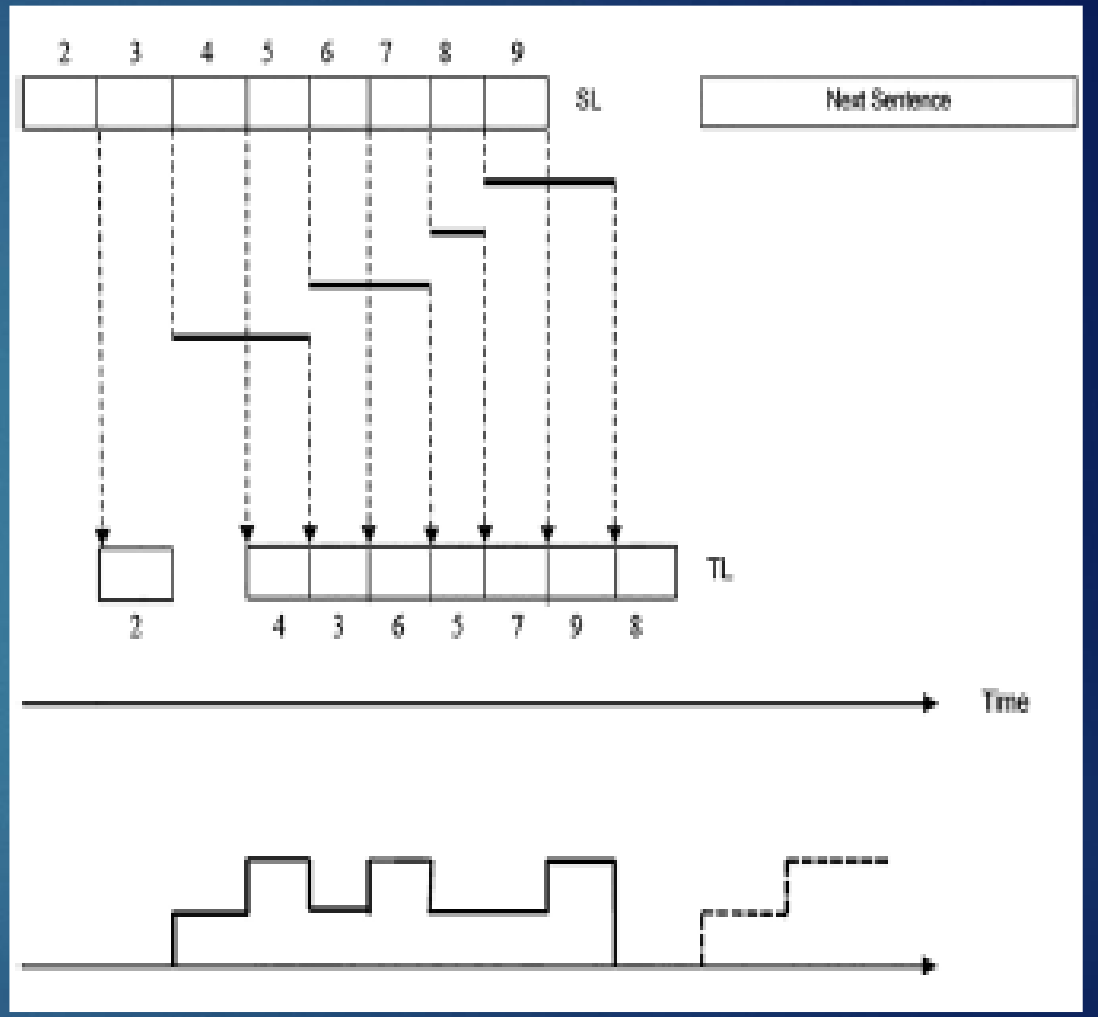


Fig.4 Translation to seek syntactic correspondence and its load
The dotted line of lower right indicates assumed load when next sentence comes in before the completion of translation of previous sentence.

Translation

Necessary #Chunk＞３！



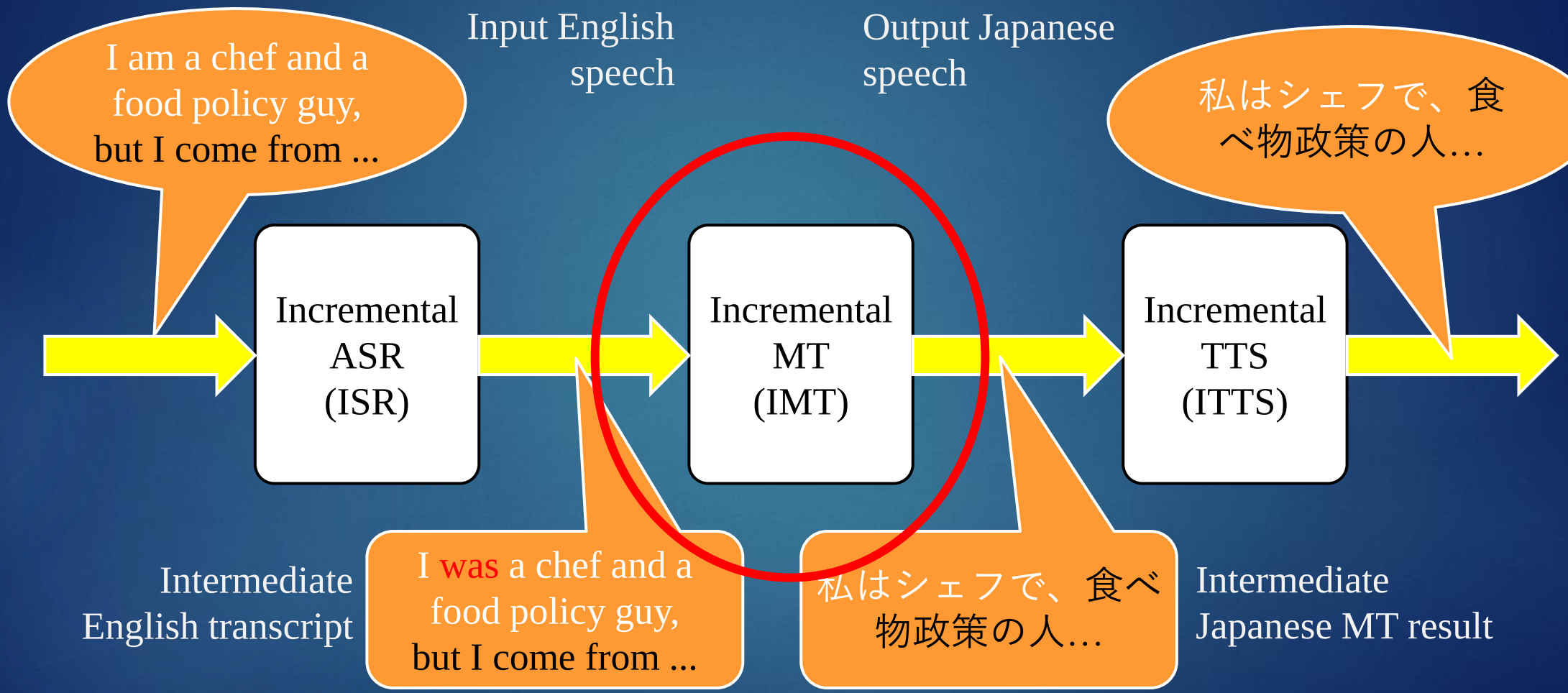
Interpretation

Necessary #Chunk＜３！

Today's contents

- Background
- Cascaded Simultaneous Speech-to-speech Translation (IWSLT2022)
- End-to-end ST and incremental TTS Speech-to-speech Translation (IWSLT2023)
- NAIST Simultaneous Speech Interpretation Corpus
- Recent Progress
  - Paralinguistic Conversion in STST
  - End-to-end Speech-to-speech Translation (TRANSLATOTRON 3)
- Summary

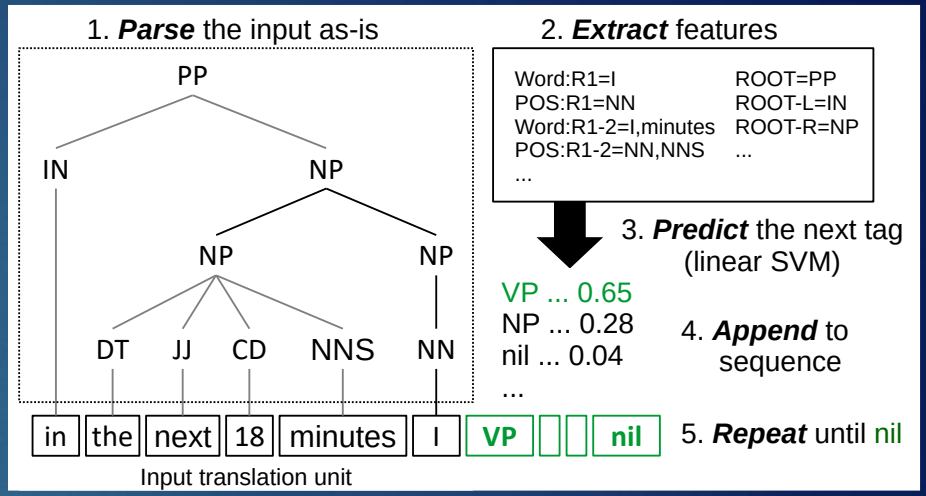
Cascade Simultaneous S2S Translation System



Translation Timing Control by Syntactic Prediction in SMT (2015,2021)

- Syntactic Prediction [Oda, et al., 2015]
  - Incremental bottom up parsing
  - Feature extraction and syntactic prediction

Subject Verb Object : English
Subject Object Verb: Japanese



- Wait for MT output when specific labels appear.
  - Control MT output timing according reordering

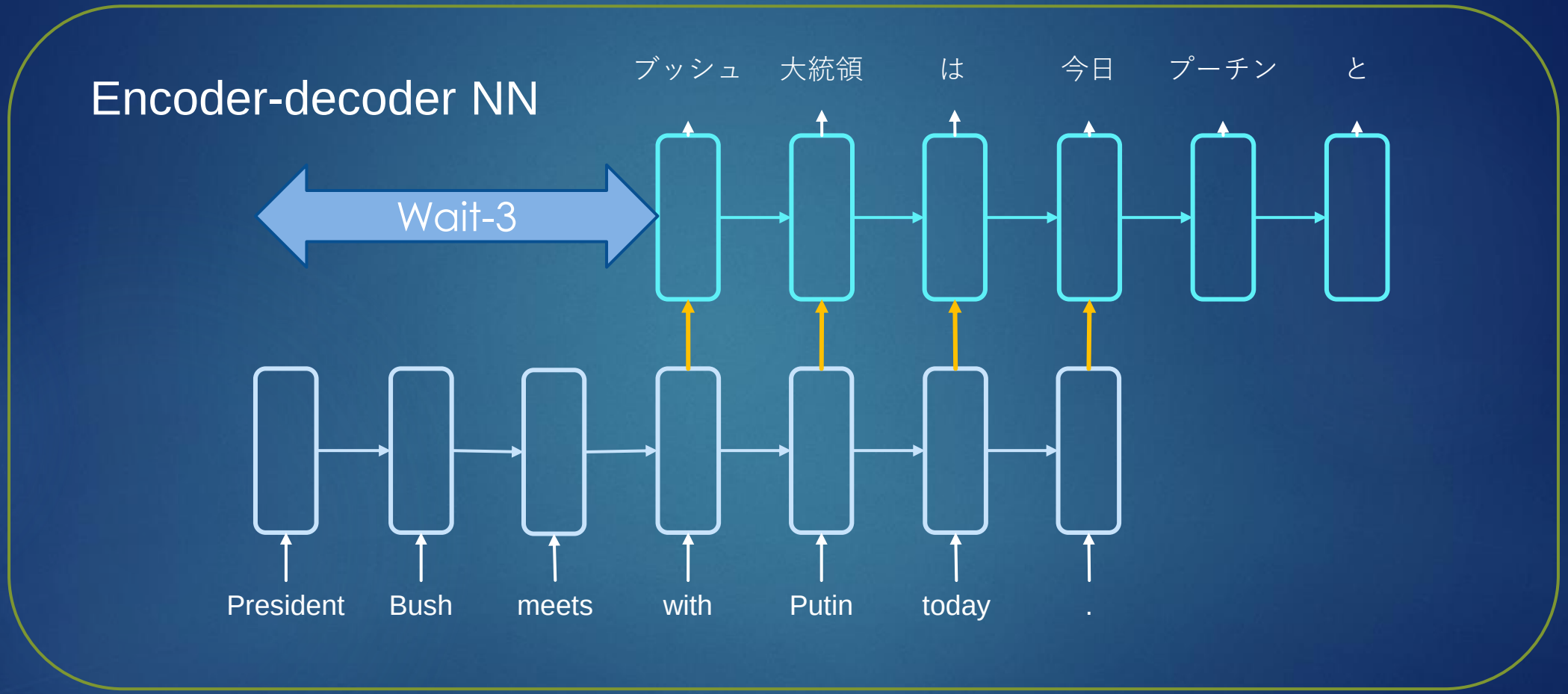
- Use LSTM and BERT to predict next tag. [Kano, et al., 2021]

‘m going to take you on a journey

Incremental parsing and syntactic prediction	in the next 18 minutes[NP] predict [VP] (wait) i 'm going to take **[keep]** i 'm going to take you on a journey [VP end]
MT results	18 分 で あ る[NP] [VP](wait) を 行 っ て い ま す[keep] 皆さん を 旅 に お連れ します [VP end]

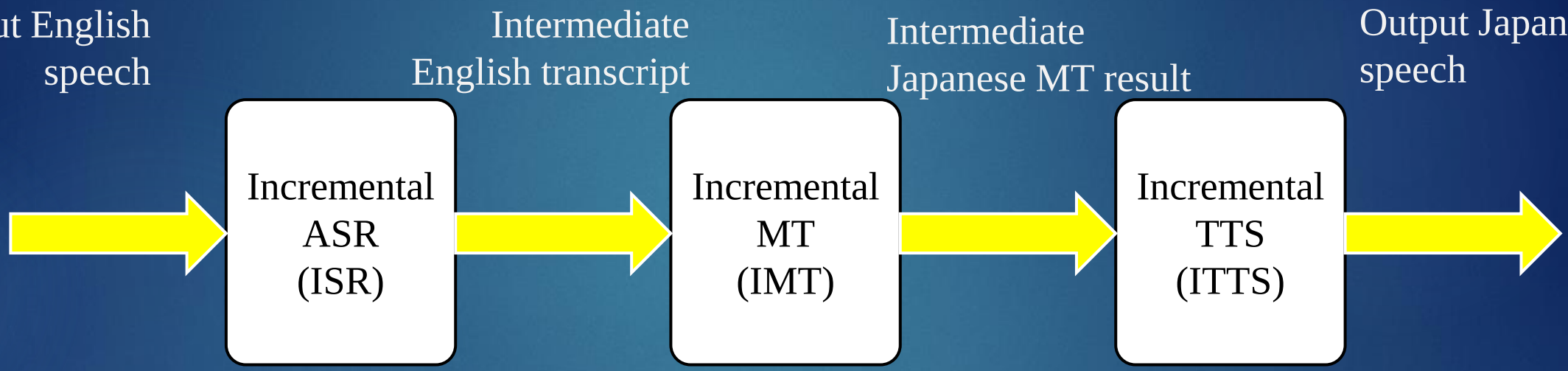
Oda, Yusuke *et al.*, Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents, Proc. of ACL-IJCNLP 2015.
Y.Kano, K.Sudo, S.Nakamura, “Simultaneous Neural Machine Translation with Constituent Label Prediction”, Proc. of WMT 2021.

Incremental MT: Wait-k Algorithm [M.Ma, et al., 2018]



Mingbo Ma, et al., “STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency”, arXiv:1810.08398v3 [cs.CL] 3 Nov 2018

NAIST IWSLT 2022 System



Incremental MT by Boundary Prediction

	Read source words	Boundary Prediction	translation
Step 1	I	⇨ 0.9 > **0.5** ⇨	私は
Step 2	I bought	⇨ 0.2 < **0.5** ⇨	
Step 3	I bought a	⇨ 0.3 < **0.5** ⇨	
Step 4	I bought a pen	⇨ 0.7 > **0.5** ⇨	私はペンを買った ↑ Forced decoding
Step 5	I bought a pen .	⇨ 0.7 > **0.5** ⇨	私はペンを買った。

The other translation Timing Control by Syntactic Prediction [Kano, et al, 2021]

ICLP: Incremental Constituent Label Prediction (LSTM, BERT)

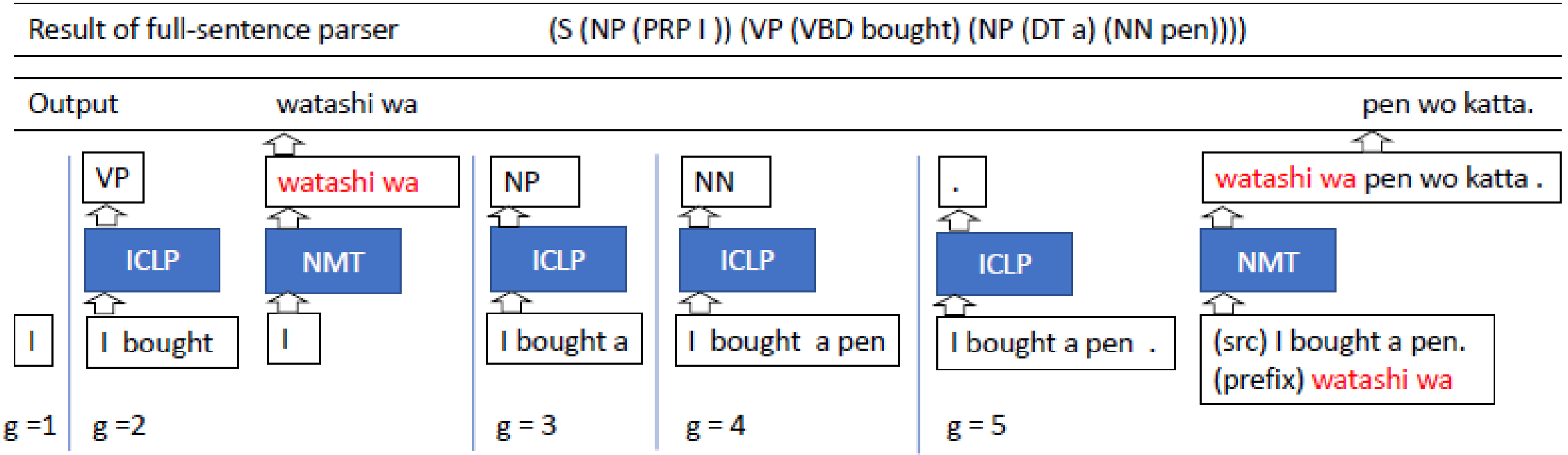


Figure 1: One look-ahead ICLP gives constituent labels. When a boundary is detected based on the label and rules, NMT starts to translate the source subsequence. The previous translation, which is red in the figure, is used as prefix words for the next translation. EOS is omitted for simplicity in the figure.

Y.Kano, K.Sudo, S.Nakamura, "Simultaneous Neural Machine Translation with Constituent Label Prediction", Proc. of WMT 2021.

Incremental Speech Recognition (ISR) [Novitasari, 2020]

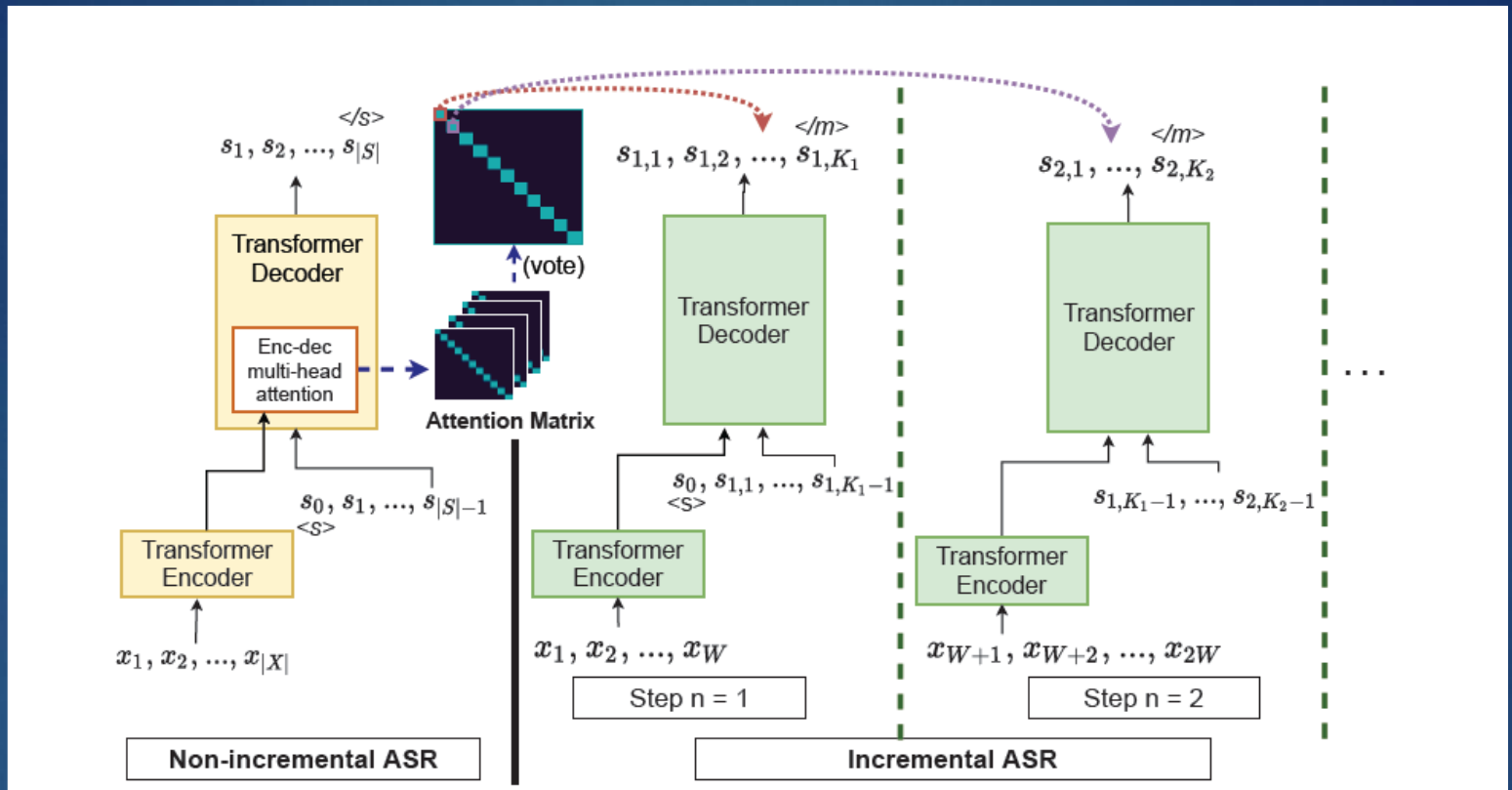


Fig. 1. Transformer-based ISR construction with attention transfer [9] from a standard Transformer-based ASR

S.Novitasari, A.Tjandra, T.Yanagita, S.Sakti, S.Nakamura, "Incremental Machine Speech Chain Towards Enabling Listening while Speaking in Real-time", Proceedings of INTERSPEECH 2020, Oct. 2020

Incremental Speech Synthesis (iTTS)

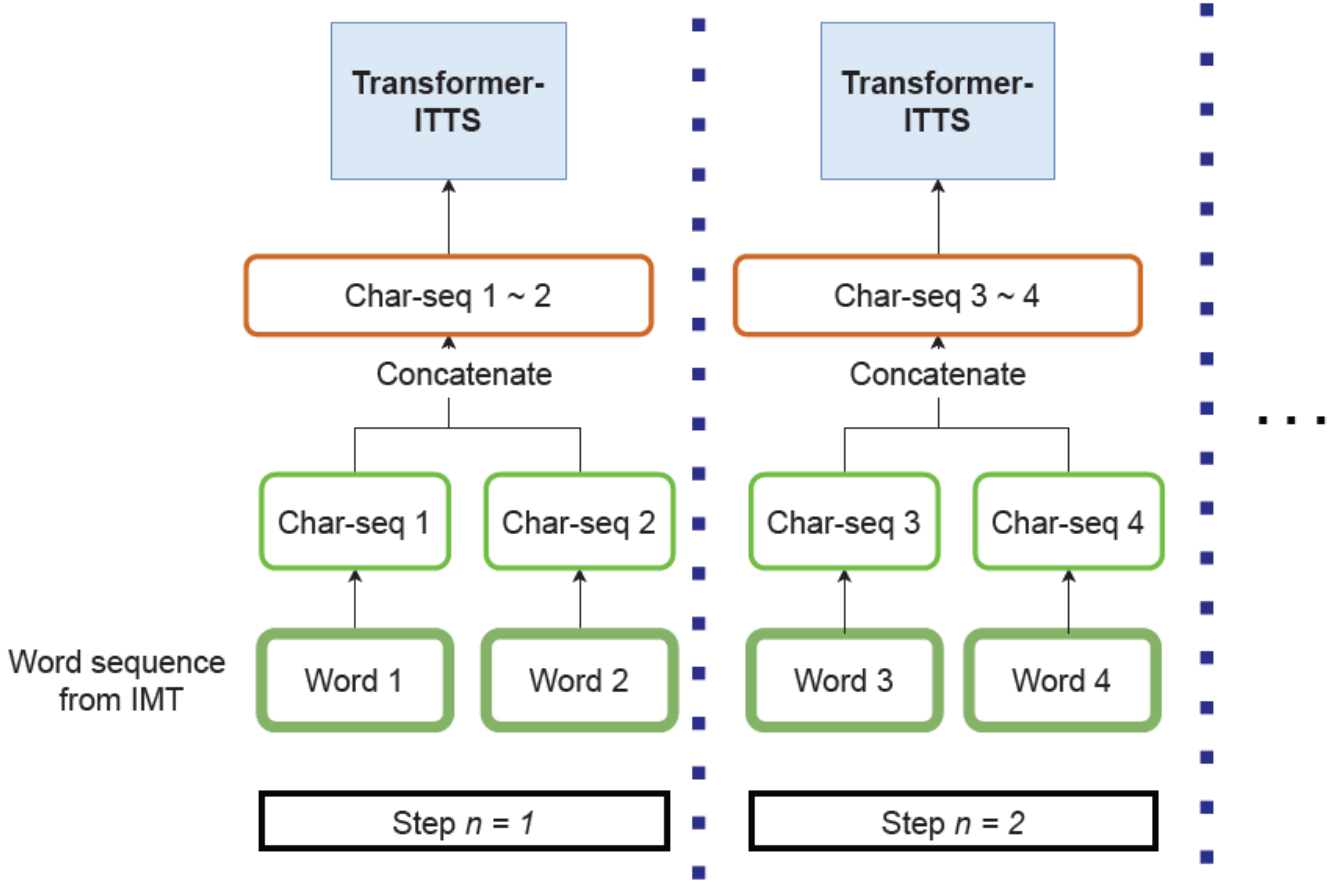


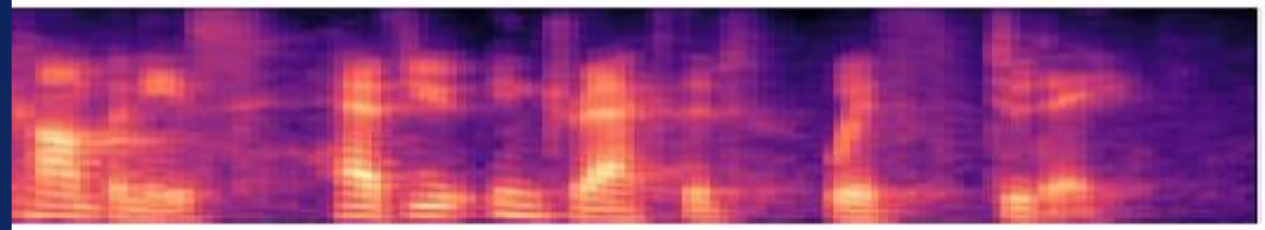
Fig. 2. Incremental text-to-speech synthesis system

S.Novitasari, S.Sakti, S.Nakamura, "Dynamically Adaptive Machine Speech Chain Inference for TTS in Noisy Environment: Listen and Speak Louder", Proc. Interspeech 2021, 4124-4128, Aug. 30, 2021

Video

How it works

Source speech spectrogram (English)



ISR & IMT &
ITTS feature synthesis
Done

1.2 how many 1.6 companies 2 have you Dore kurai 2.4 interact no 2.8 ed Kaisha 3.2 with ga 3.6 today yaritori 4 </s> 5.6

wo | shiteiru ka shimeshi te iru |

kaisha-ga-yaritori |wo shite iruka shimeshi teiru |

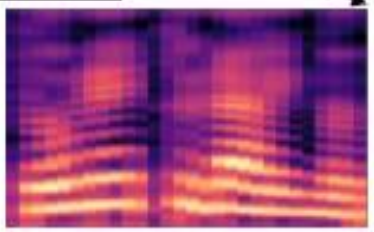
Dore-kurai-no

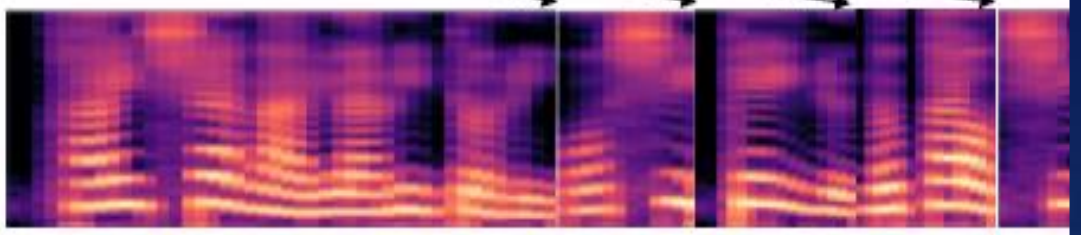
ISR delay

IMT delay

ITTS delay

delay

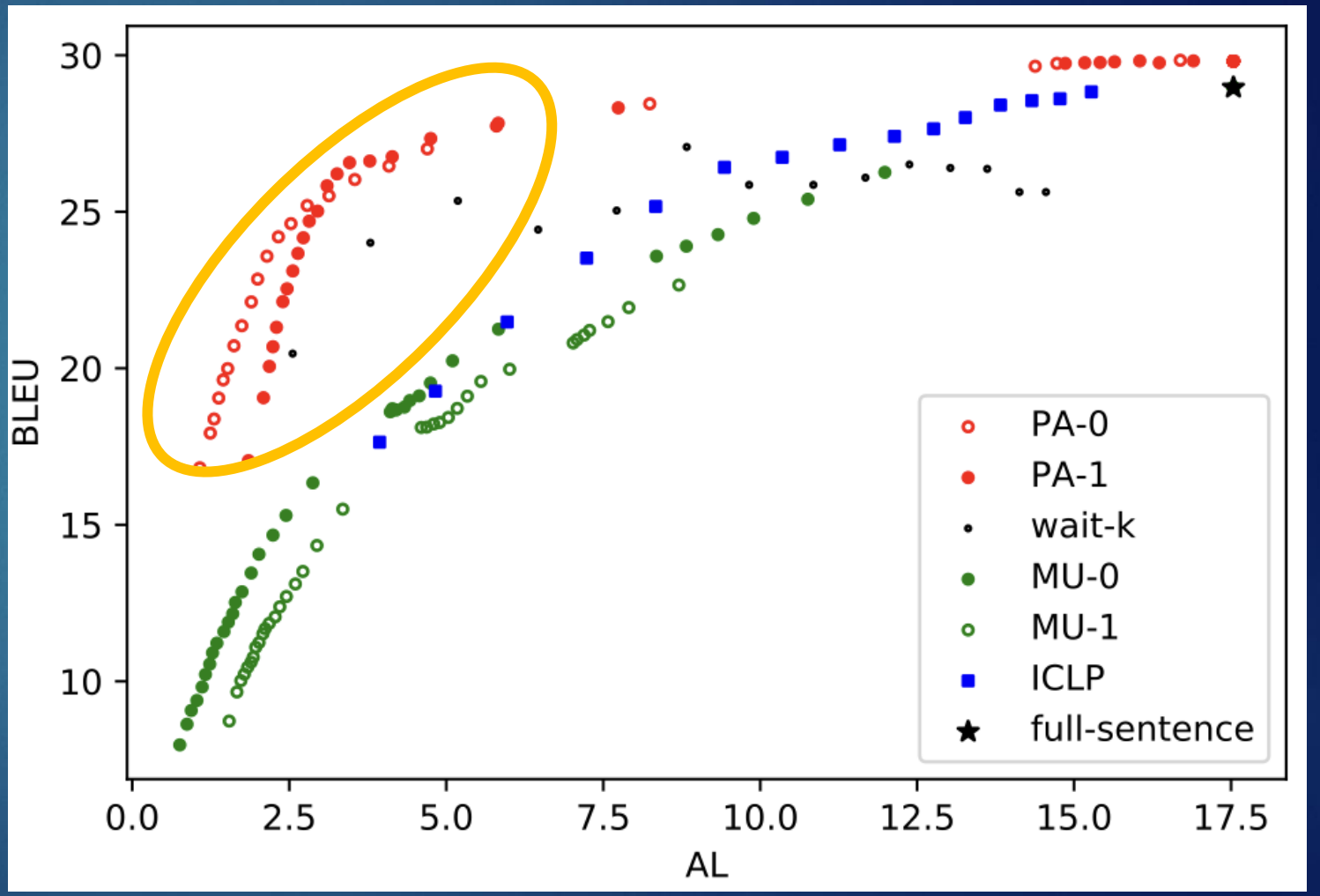




Target speech spectrogram(Japanese)

Results

- SimulMT model trained using the prefix pairs outperformed other methods (En-De)
- The advantage is smaller in En-Ja; PA failed to induce enough short prefix pairs that helps SimulMT



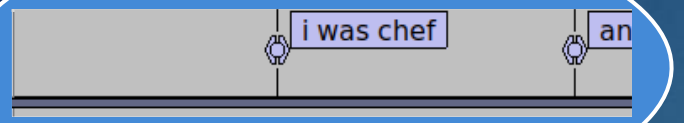
Today's contents

- Background
- Cascaded Simultaneous Speech-to-speech Translation (IWSLT2022)
- End-to-end ST and incremental TTS Speech-to-speech Translation (IWSLT2023)
- NAIST Simultaneous Speech Interpretation Corpus
- Recent Progress
  - Paralinguistic Conversion in STST
  - End-to-end Speech-to-speech Translation (TRANSLATOTRON 3)
- Summary

2023 SimulS2S System [Fukuda+ 2023 IWSLT]

Input English speech

Output Japanese speech



Incremental ST (IST)

Incremental TTS (ITTS)

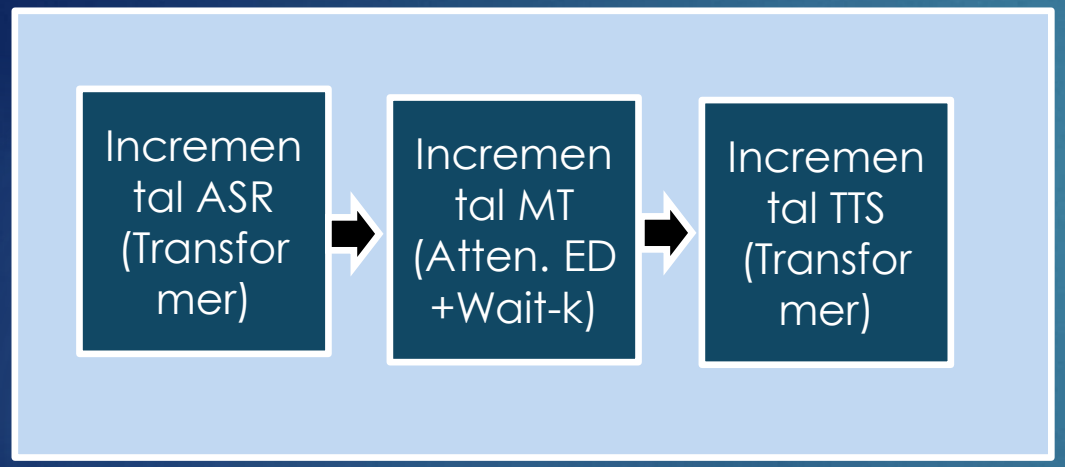
No explicit English transcriptions

私はシェフで、食べ物政策の人…

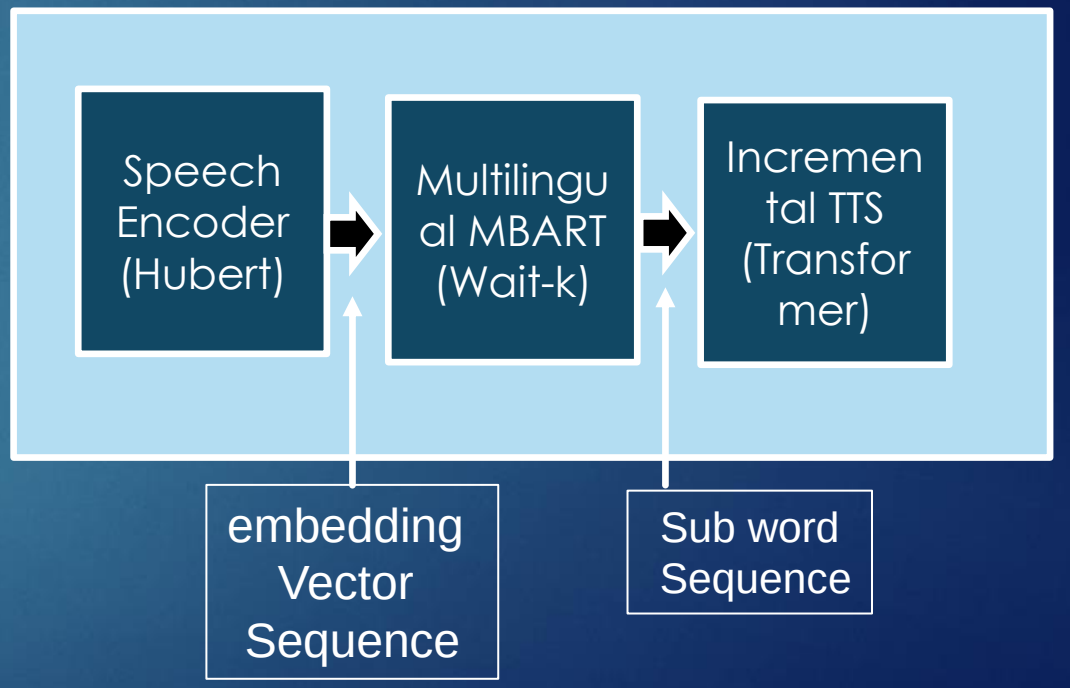
Intermediate Japanese MT result

IWSLT NAIST 2023 SYSTEM

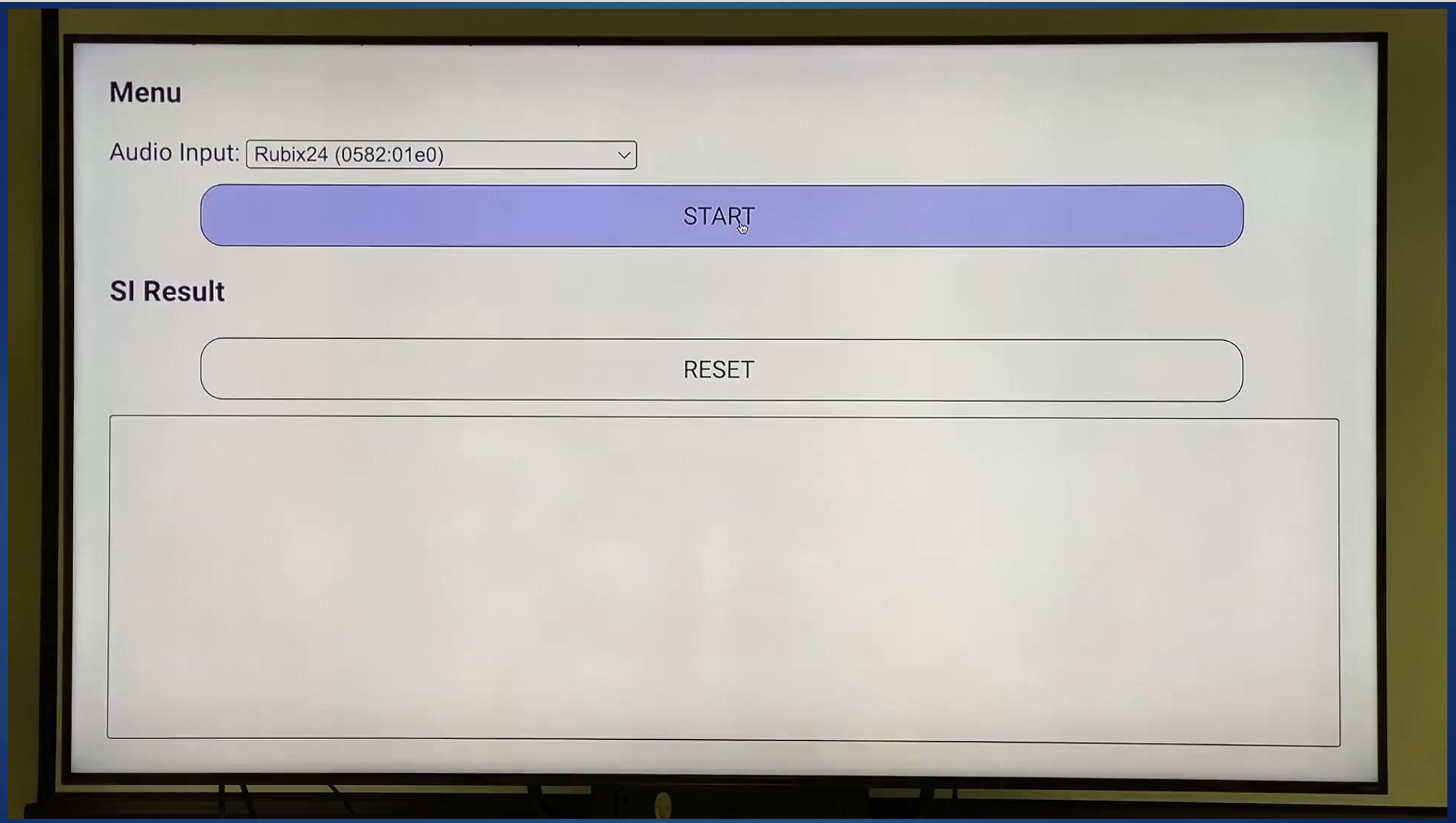
IWSLT 2022 STST System



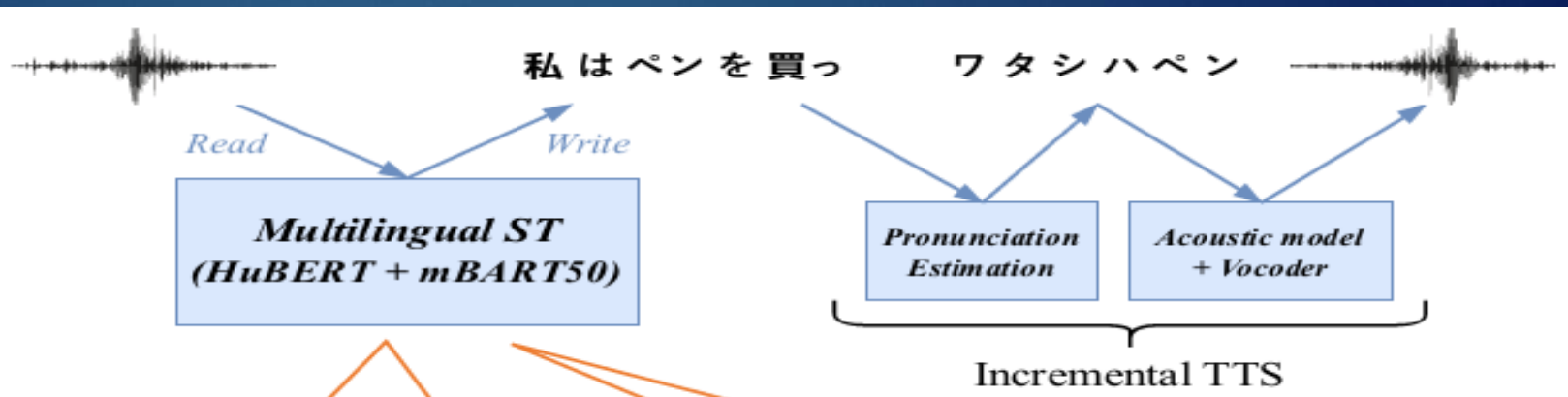
IWSLT 2023 STST System



Video (En-Ja, speech-to-speech)



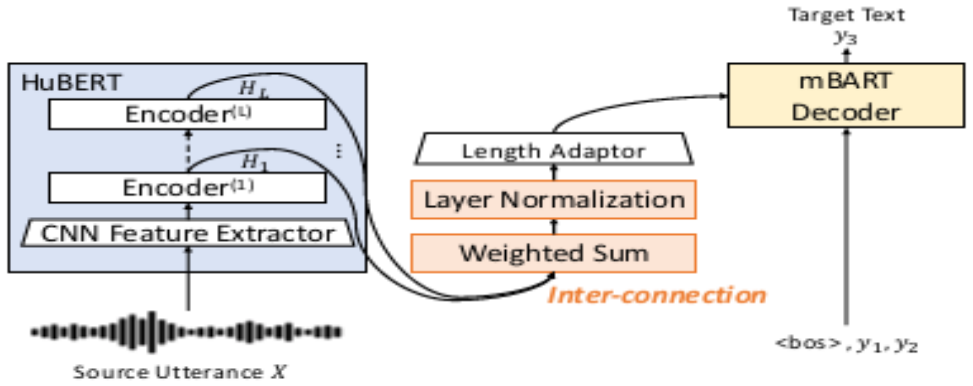
System Structure



Inter-connection [Nishikawa+INTERSPEECH2023]

Effective connection between pre-trained Encoder and Decoder

- Aggregate hidden states from intermediate layers of HuBERT
  → Input it to the mBART Decoder



Prefix Alignment [Kano+IWSLT2022]

Data augmentation for prefix-to-prefix translation

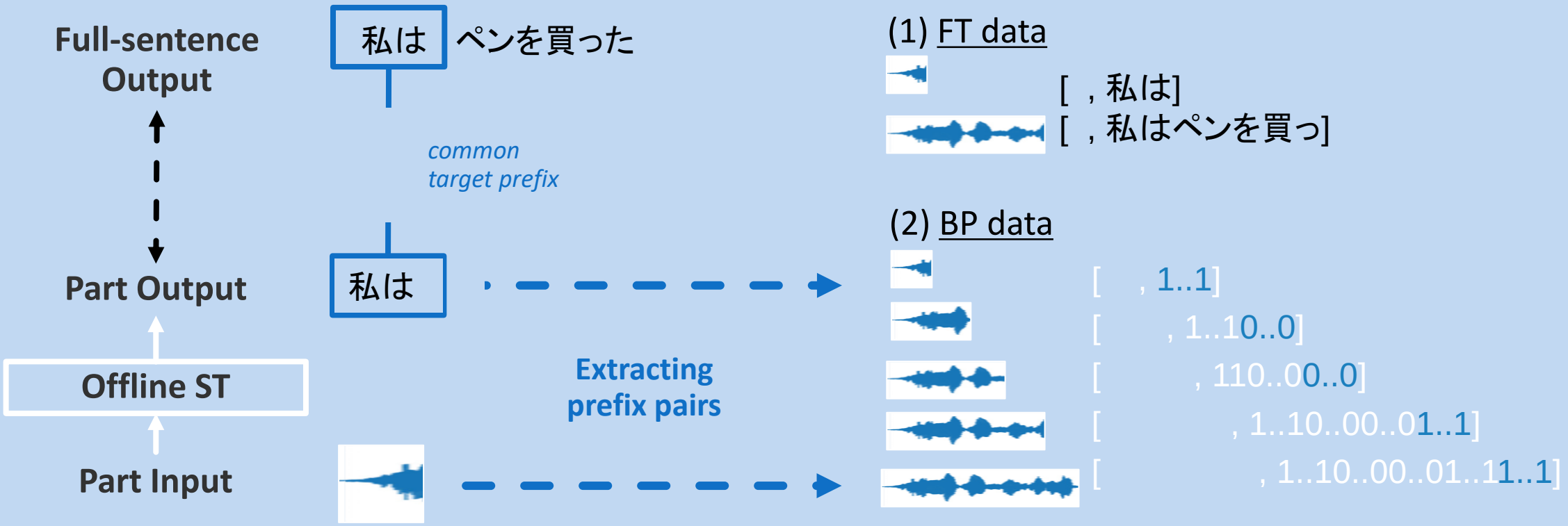
- Extract prefix pairs using offline translation
  → fine-tune an offline ST model for SimulST

Source Prefix	Prefix Translation (*gloss*)	Offline translation
I	私は。 (*I*.)	私はペンを買った
I bought	私は買った。 (*I bought*)	
I bought a	私は一つ買った (*I bought one*)	
I bought a pen	私はペンを買った (*I bought a pen*)	

Prefix pairs

[("I", "私は"),
("I bought pens.", "私はペンを買った。")]

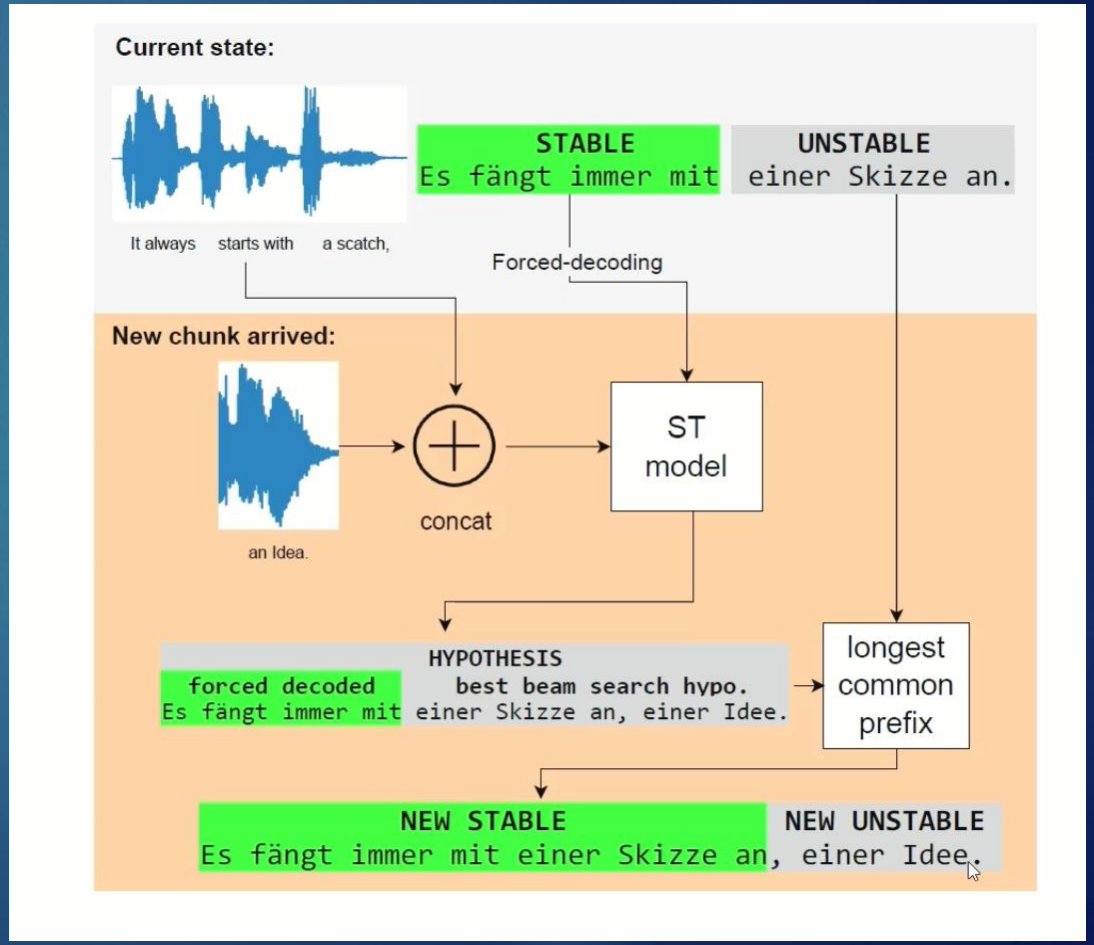
Fine-tuning by Prefix Alignment [Kano+ 2022 IWSLT]



Local Agreement: Chunk-wise Inference with *stable hypothesis*

Incremental decoding described by Liu et al., (2020)

1. Audio is split to chunks
2. After the system receives a new chunk, it starts with forced-decoding of the *stable hypothesis* and then continues with beam search



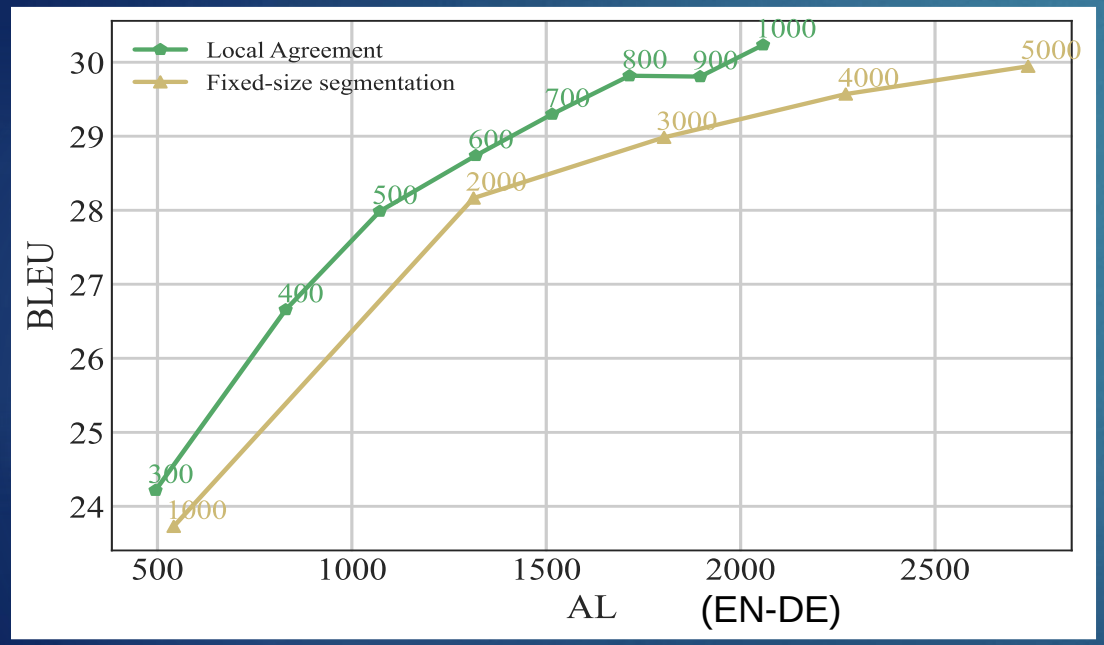
Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In Proc. Interspeech 2020.

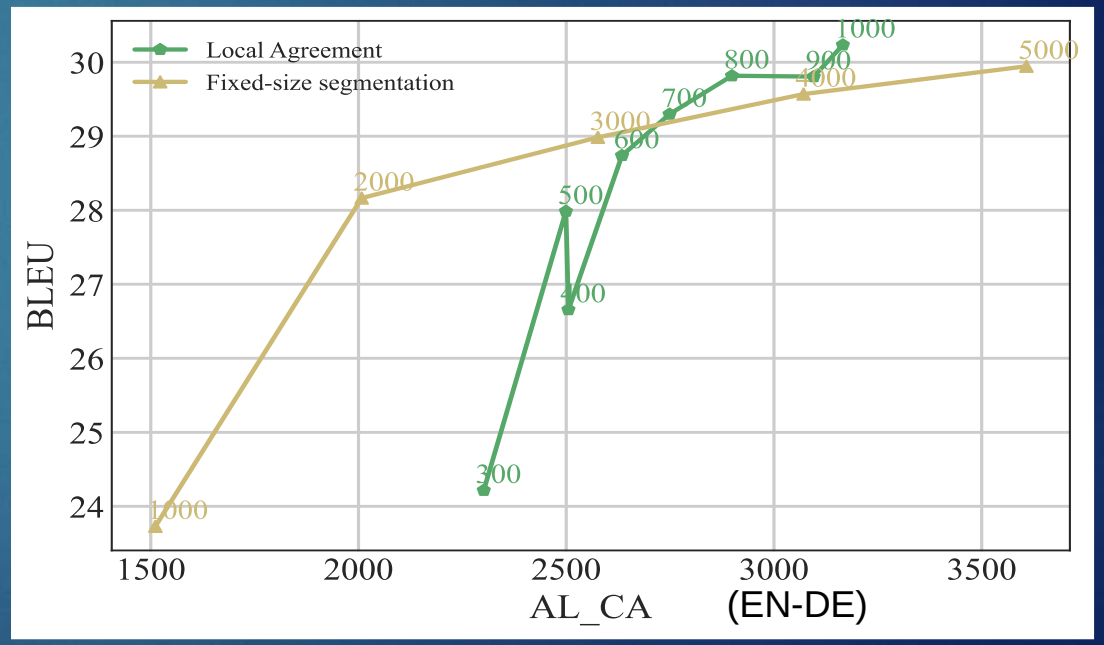
Results

Lang pair	chunk size	BLEU	AL
En-De	950 ms	29.98	1964
En-Ja	840 ms	15.32	1974
En-Zh	700 ms	22.11	1471

Results with and w/o Inter-connection.

Model	En-De	En-Ja	En-Zh	Ave.
HuBERT + mBART	30.47	15.71	**25.01**	23.73
w/ Inter-connection	**30.89**	**15.89**	24.75	**23.84**





Computation-aware AL

IWSLT Evaluation Campaign – SimulST track

- History
  - 2020: text/speech-to-text, En-De
  - 2021-2022: text/speech-to-text, En-De/**Ja**/**Zh**
  - 2023: speech-to-text/**speech**, En-De/Ja/Zh
  - We're planning the next edition (2024)
- Regulations
  - Use publicly-available speech/language resources
  - Configure SimulST systems to satisfy given latency limits
  - Submit systems in form of Docker images

Today's contents

- Background
- Cascaded Simultaneous Speech-to-speech Translation (IWSLT2022)
- End-to-end ST and incremental TTS Speech-to-speech Translation (IWSLT2023)
- NAIST Simultaneous Speech Interpretation Corpus
- Recent Progress
  - Paralinguistic Conversion in STST
  - End-to-end Speech-to-speech Translation (TRANSLATOTRON 3)
- Summary

NAIST-SIC (Simultaneous Interpretation Corpus)

- A collection of Simultaneous Interpretation
  - https://dsc-nlp.naist.jp/data/NAIST-SIC/
  - You can (easily) find by searching “NAIST-SIC”
  - A part of this corpus (NAIST-SIC 2021) was used for IWSLT Simutaneous Translation shared task
    - It was also presented at IWSLT 2021
      - Doi et al., Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data, Proc. IWSLT 2021.
  - An additional release (NAIST-SIC 2022) includes automatic source-target sentence alignment

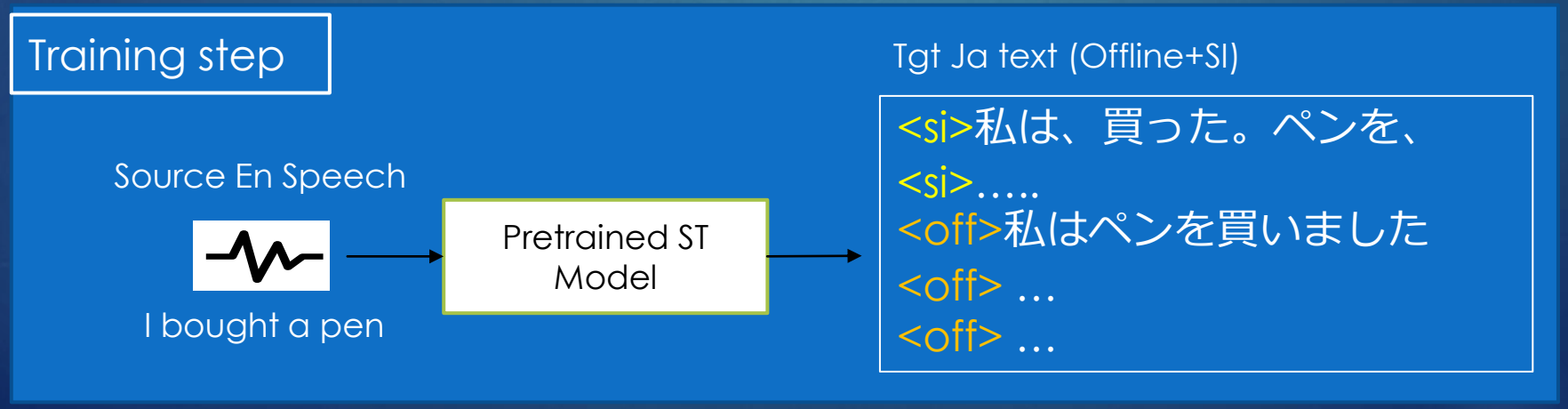
Corpus Development

- We collected SI recordings ~10 years in our government-funded project and released a part of them:
  - NAIST-SIC 2014: En-Ja/Ja-En
  - NAIST-SIC 2021: En-Ja (TED Talks), 57 SI transcripts on 28 talks
  - NAIST-SIC 2022: En-Ja (TED Talks), 833 SI transcripts on 831 talks with automatic sentence alignment
- We’re planning to release remaining portions (esp. Ja-En) by March 2024
  - Our recordings reach 340 hours in total

Using SI Corpus for SimulST

- A naïve use of the corpus was not so effective
  - Reported in our preprint: NAIST-SIC-Aligned (arXiv: 2304.11766)
- Mixed use of offline translation and SI (Ko+ 2023 IWSLT)
  - Introducing style tags <off> <si> to distinguish data sources
  - Used to fine-tune pre-trained ST models

	BLEU	
	SI ref.	*Offline* ref.
Offline fine-tune	7.8	16.0
SI fine-tune	10.9	6.3
Mixed fine-tune	9.4	13.3
Proposed	10.3	15.4
Proposed + up-sampling SI data	12.2	14.2

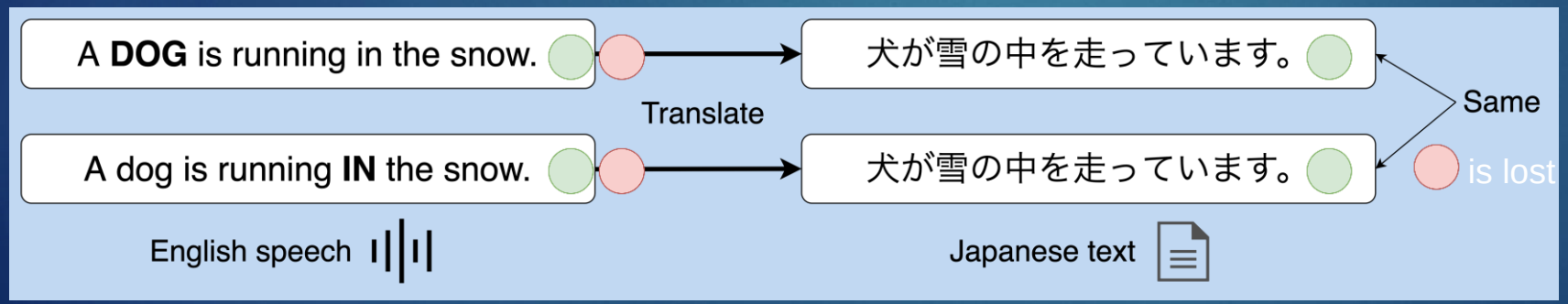


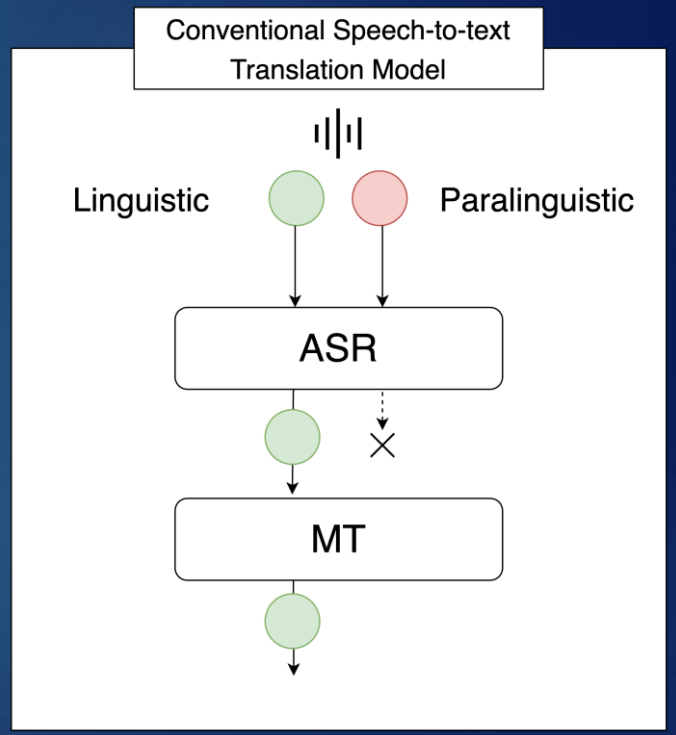
Today's contents

- Background
- Cascaded Simultaneous Speech-to-speech Translation (IWSLT2022)
- End-to-end ST and incremental TTS Speech-to-speech Translation (IWSLT2023)
- NAIST Simultaneous Speech Interpretation Corpus
- Recent Progress
  - End-to-end Speech-to-speech Translation (TRANSLATOTRON 3)
  - Paralinguistic Conversion in STST
- Summary

Speech Translation and its Limitation

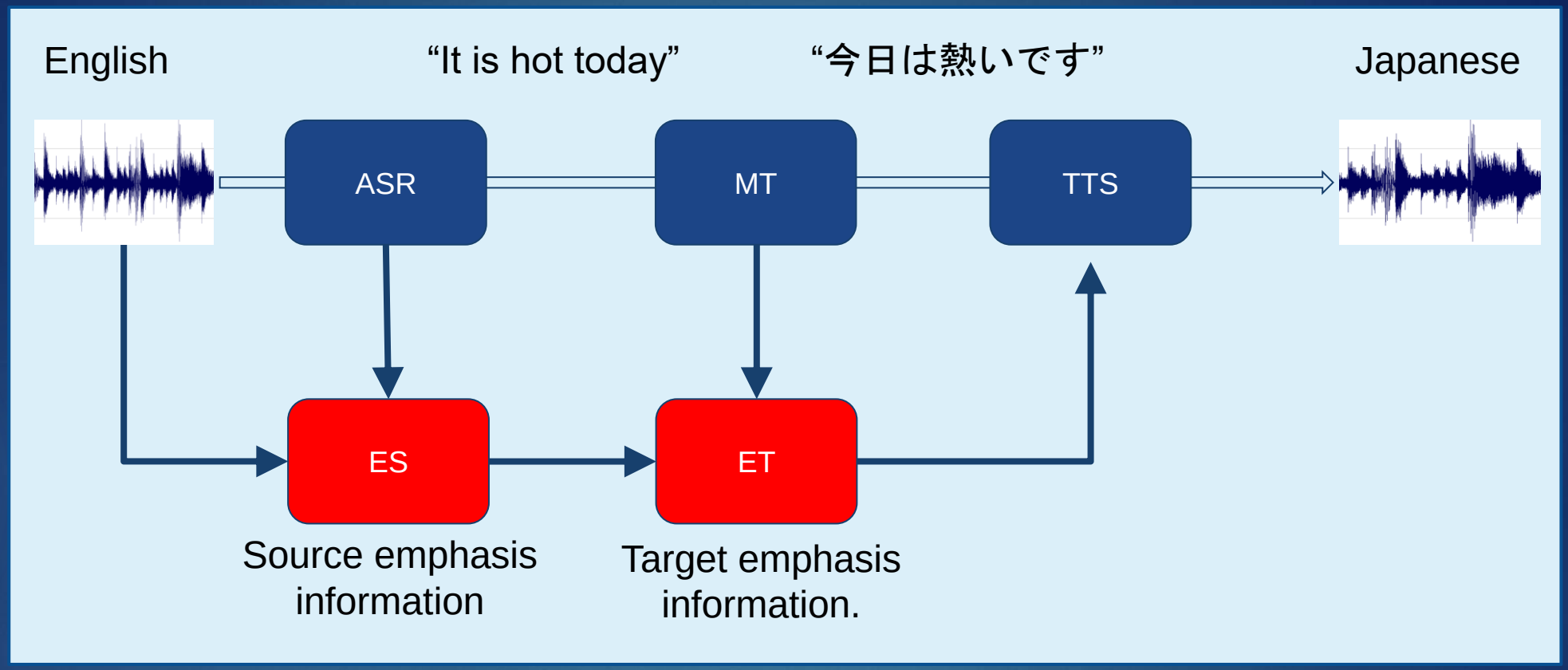
- What is it?
  - Translates speech in one language into text/speech in another
- Limitation
  - Unable to consider paralinguistic info
  - If the linguistic info is the same, so are the translations.





- Could cause misunderstandings between a speaker and a hearer

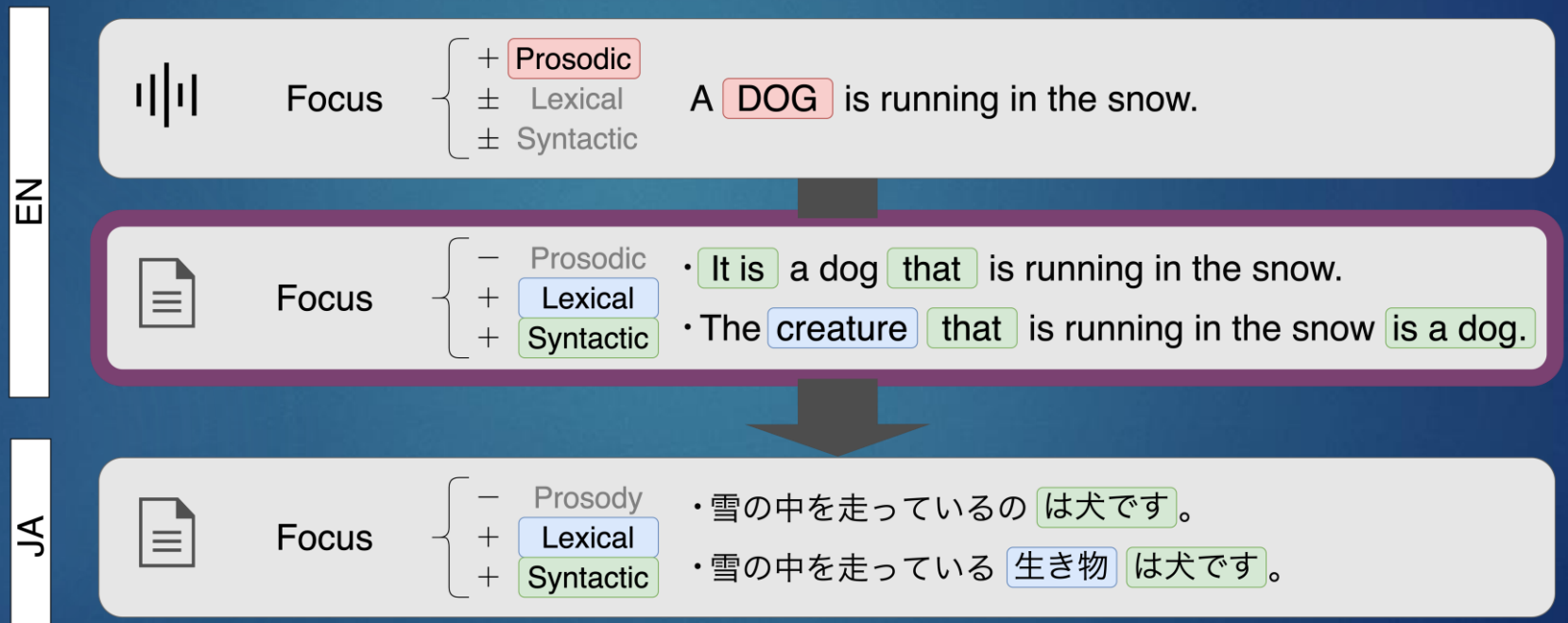
Paralinguistic Information in S2ST



(1) Emphasis estimation (ES) systems: by speech and word sequence
(2) Emphasis translation (ET) systems: by mapping emphasis to the target language.

Paralinguistics Conversion II

- Paraphrase focused speech into focused text
  - Mapping Prosodic Lexical it Syntactic
  - Translating the paraphrased text into the TL



- Could return different translations depending on the focused items

N.Suzuki, S.Nakamura, “Representing ‘how you say’ with ‘what you say’: English corpus of focused speech and text reflecting corresponding implications”, INTERSPEECH2022

Summary

- Recent Speech-to-speech Translation Systems
  - Cascaded Simultaneous Speech-to-speech Translation (IWSLT2022)
  - End-to-end ST and incremental TTS Speech-to-speech Translation (IWSLT2023)
- NAIST Simultaneous Speech Interpretation Corpus
- Recent Progress
  - Paralinguistic Conversion in STST

Future Directions

- Analysis of the interpretation data
- More practical and efficient Simul-SST
  - Beyond the use of offline translation data
  - Effective use of the interpretation data
- Sophisticated quality and latency assessment for Simul-SST
  - Applicable to SI and human interpretation
- Pretrained LLM and fine-tuning
- Summarization for text and output speech
- Paralinguistic/nonlinguistic information in Simul-SST

Thank you for your attention!