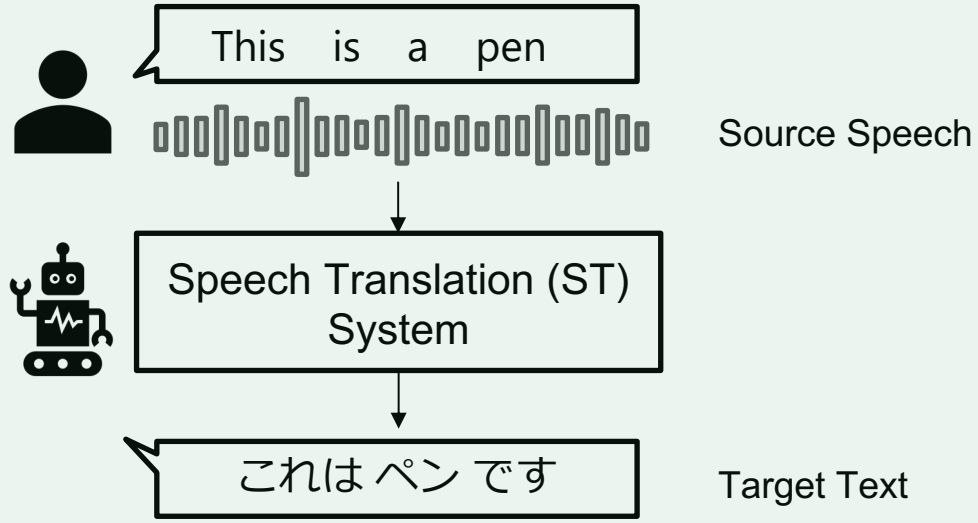
# Optimizing Speech Translation for Low Latency and High Robustness

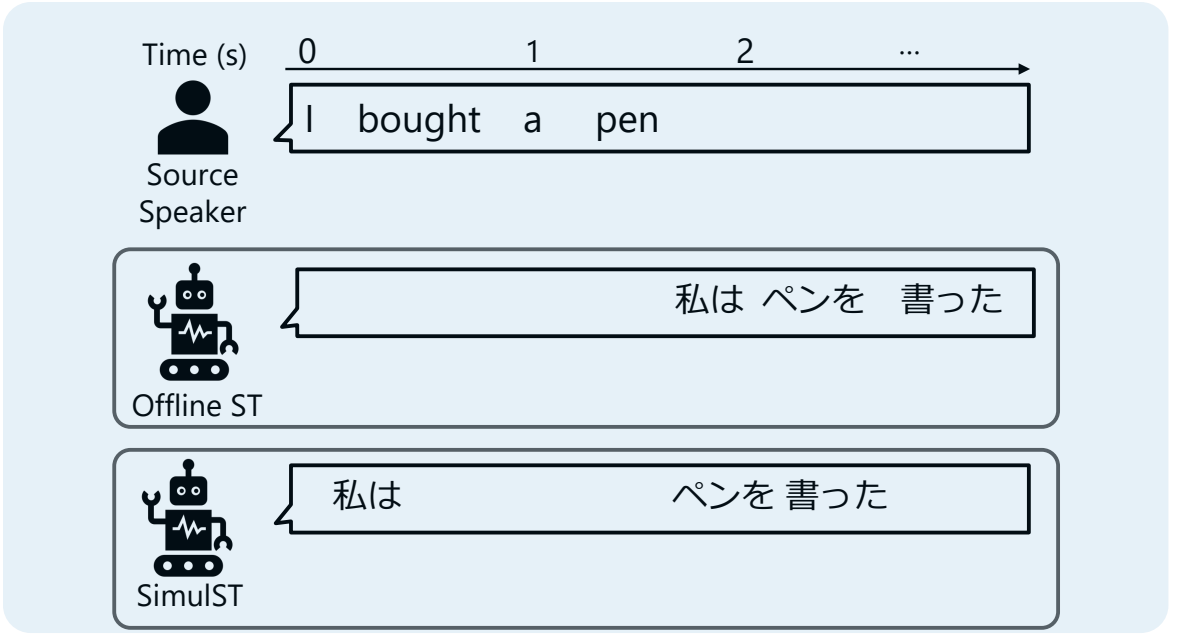
胡　尤佳
奈良先端科学技術大学院大学 ヒューマンAIインタラクション研究室

2025.06.18
第12回AAMT長尾賞学生奨励賞招待講演

# 概要：本研究が目指すもの

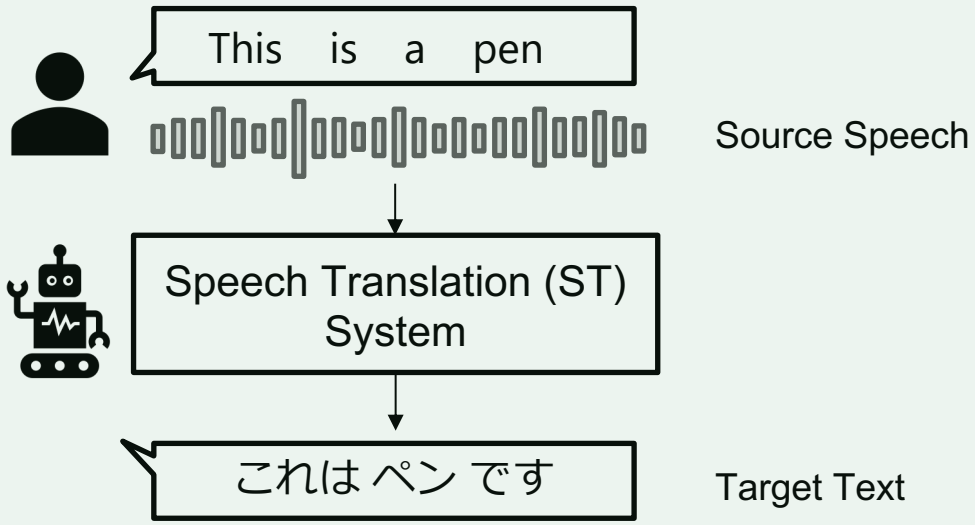
- 現実のシチュエーションで使える高頑健かつ低遅延な音声翻訳システム
  - 以下に取り組んだ
    - (A) 話し言葉に含まれる曖昧な入力に強いEnd-to-end音声翻訳
    - (B) 同時通訳データを用いたEnd-to-end同時音声翻訳と学習手法

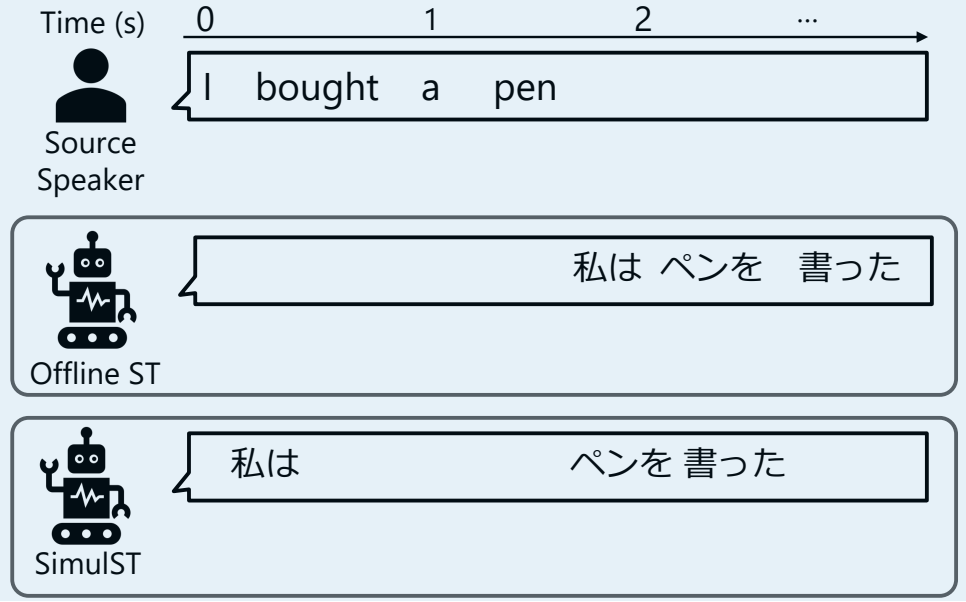


# 背景

- 音声翻訳 (Speech Translation; ST)
  - 原言語 → 目的言語テキストや音声 (本研究：テキスト出力)
  - 一般的に文単位のオフライン音声翻訳を指す (Offline Speech Translation)
- 同時音声翻訳 (Simultaneous Speech Translation; SimulST)
  - 話し手が話し終わるのを待たずに翻訳できるタイミングで訳出を進める

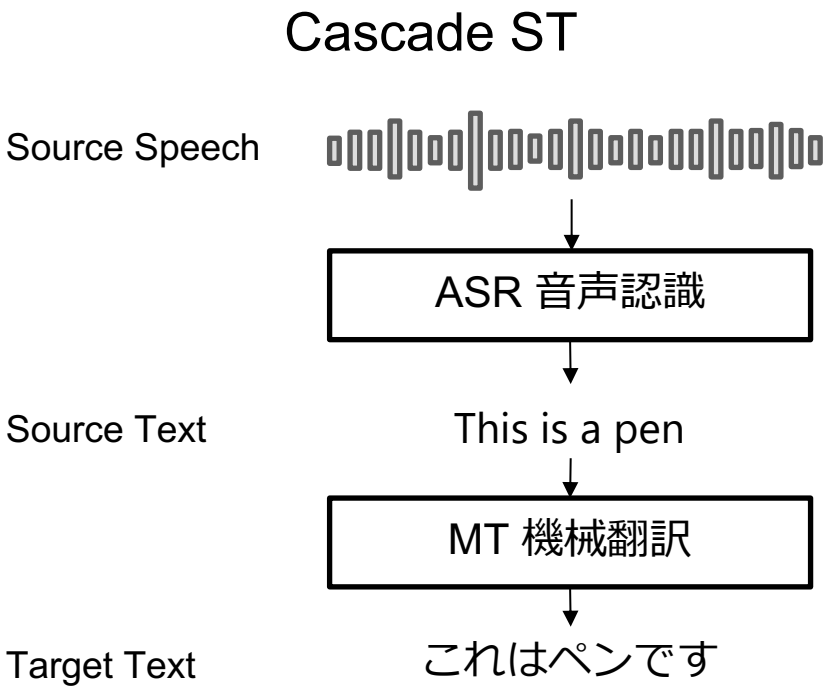


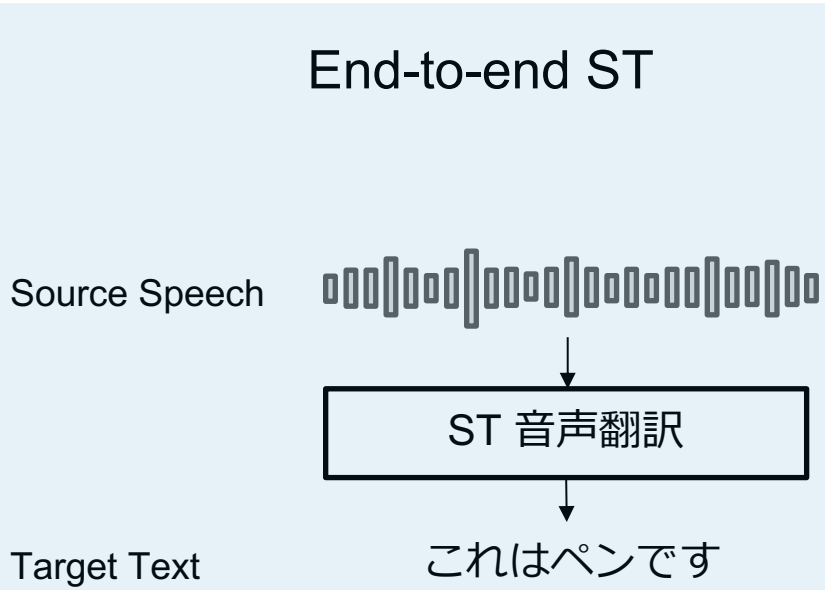
# 近年の音声翻訳 (ST)

■ ニューラルベースモデル：学習のシンプルさからEnd-to-endが最近では主流

- 本研究でもEnd-to-end STを採用




- パラメータが多く全体のチューニングが難しい



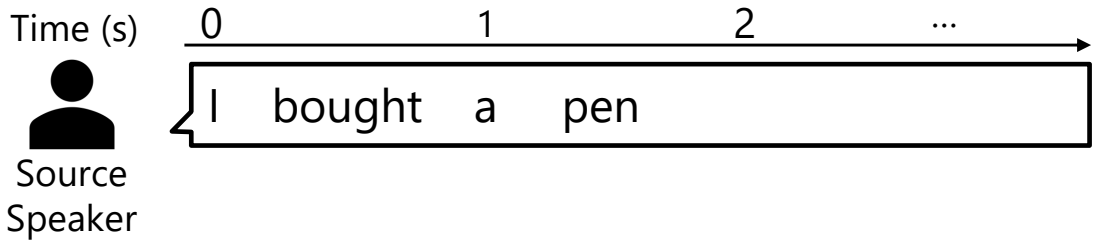

- **より少ないパラメータでLow latencyなモデルを実現**

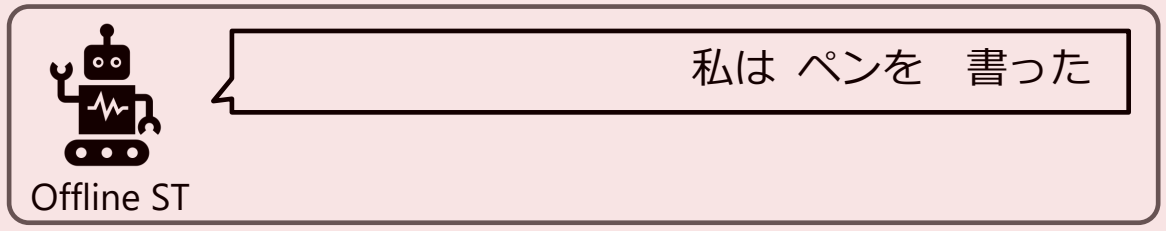
# 課題1：音声翻訳では話し言葉に含まれる曖昧性に弱い

- Prepared Speech で学習されたモデルは **Spontaneous Speech** に弱い
  - 事前に台本のない自発的な話し言葉：曖昧な発話、フィラーなどを多く含む
  - 曖昧性の少ないprepared speechによるデータがほとんど
  - 曖昧性やフィラーがあるデータは少ない → 学習の工夫が必要

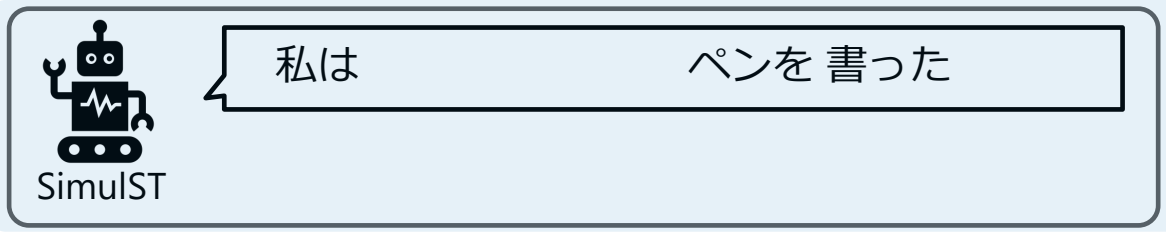
# 課題2：同時音声翻訳での同時通訳タスクとデータの乖離

- 現在の同時音声翻訳：一般的な文単位音声翻訳データ(Offlineデータ)での学習がほとんど
  - Offline ↔ **Simultaneous**
  - Prepared speech in Offline setting: **現実での同時通訳タスクとの乖離(ミスマッチ)**
- モデル学習に使える同時通訳データが少量→ 学習を工夫する必要




- タスク: 翻訳
- 原言語を全て聞いてから訳出をスタート




- タスク: 同時通訳
- 話し手が話し終わる前から訳出を開始

# 本研究が目指すもの

■ 以上の課題に対するモデルの提案

- (課題1) 話し言葉に含まれる曖昧性に弱い
  - (A) 話し言葉に含まれる曖昧な入力に強いEnd-to-end音声翻訳

- (課題2) 同時通訳タスクとデータの乖離
  - (B) 同時通訳データを用いたEnd-to-end同時音声翻訳と学習手法

■ 現実のシチュエーションで使える高頑健性かつ低遅延な音声翻訳システム

# 本研究が目指すもの

■ 以上の課題に対するモデルの提案

- (課題1) 話し言葉に含まれる曖昧性に弱い
  - ➢ (A) 話し言葉に含まれる曖昧な入力に強いEnd-to-end音声翻訳

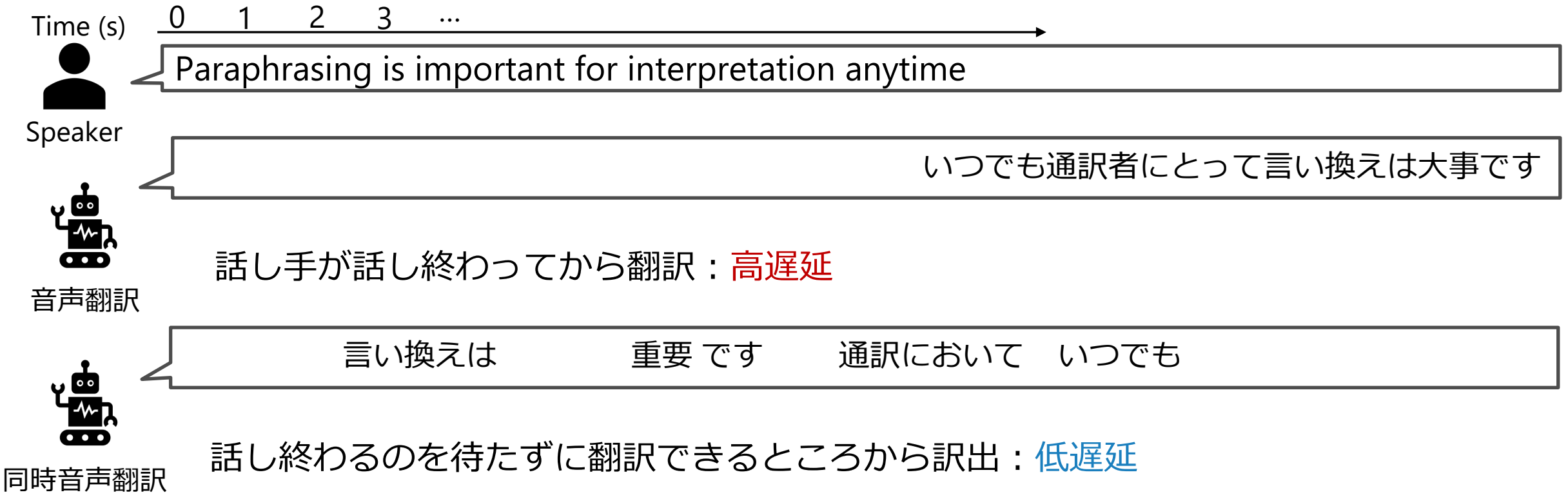
- (課題2) 同時通訳タスクとデータの乖離
  - ➢ (B) 同時通訳データを用いたEnd-to-end同時音声翻訳と学習手法

本日話す内容

■ 現実のシチュエーションで使える高頑健性かつ低遅延な音声翻訳システム

# 同時通訳データを用いたEnd-to-end同時音声翻訳と学習手法

# 同時音声翻訳 (Simultaneous Speech Translation; SimulST)




■ 同時通訳者: 入力された順に順送り訳、省略など行う → 話し手が次々に話す内容に追従可能

やりたいこと : 同時通訳データを用いたモデルの学習 → 低遅延かつ高品質な同時音声翻訳

課題 : モデル学習に使える同時通訳データが少量→ 学習を工夫する必要

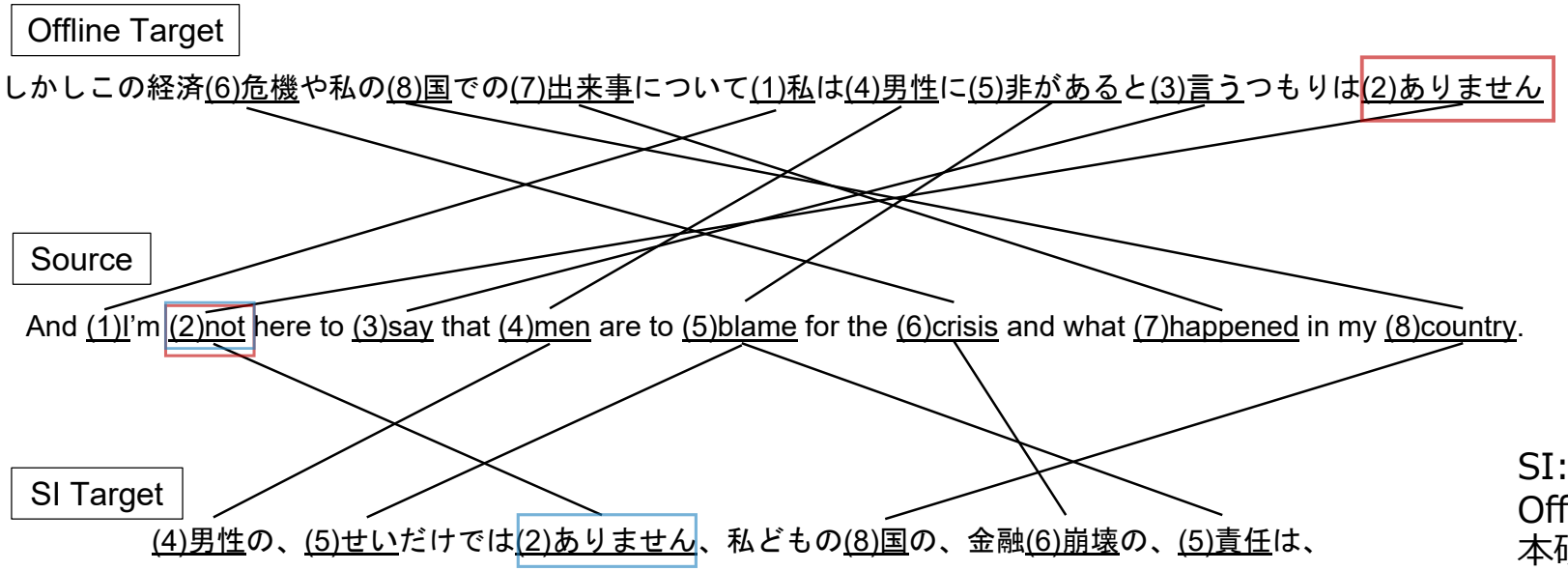
# OfflineデータとSIデータの違い

- 一般的な翻訳 (Offline)
  - 英語に対応する日本語がほとんど存在し省略がない
  - 英語-日本語で対応する単語間の距離が長い場合も

→ 入力の内容を削らない流暢な出力

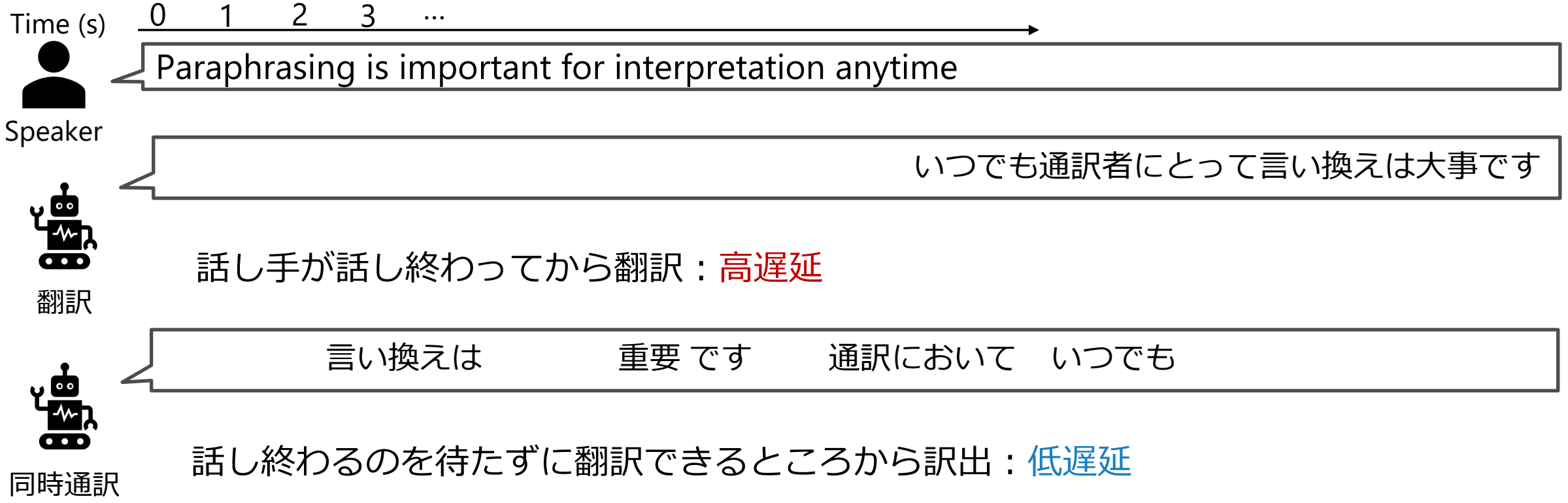
- 同時通訳 (SI ;Simultaneous Interpretation)
  - 優先度の低い部分は省略
  - 英語-日本語で対応する単語間の距離が短い

→ できるだけ早く話し手の意図を伝える
話し手の発話に追いつくことができる




SI: NAIST-SIC-Aligned-ST
Offline: MuST-C TEDより例
本研究の実験でも利用

# 同時音声翻訳 (SimulST)

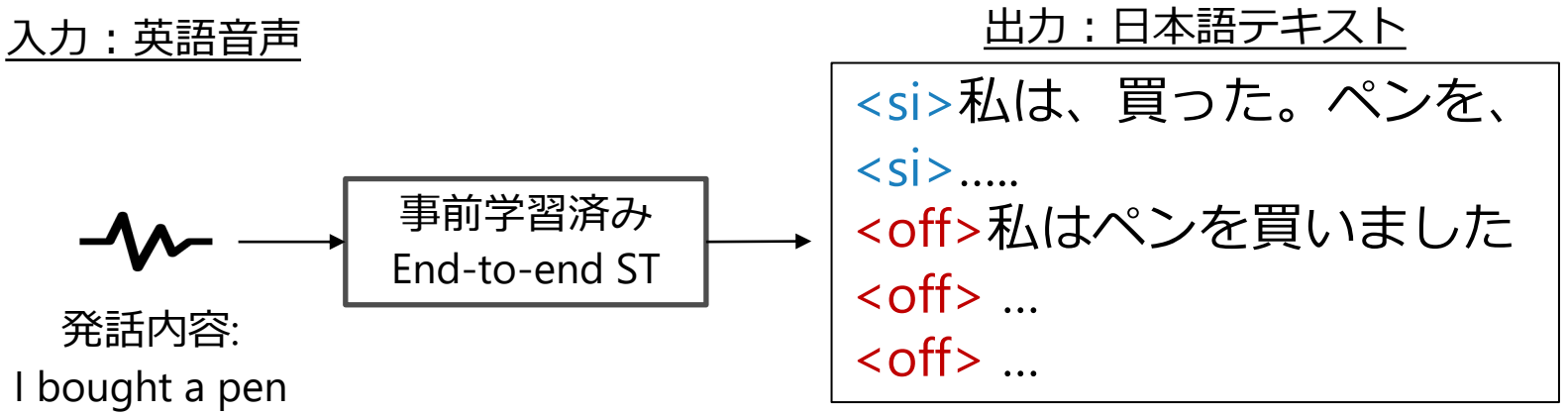



- 課題：モデル学習に使える同時通訳データが少量→ 学習を工夫する必要
- 過学習を起こしやすい：SIはOfflineよりも省略があり短い→それを学習して出力を削り過ぎる傾向

  → 少量のSIデータだけでなく、比較的多量用意できるOfflineデータも活用したい

# 提案法
## スタイルタグ付きデータでの混合学習 (Style FT)

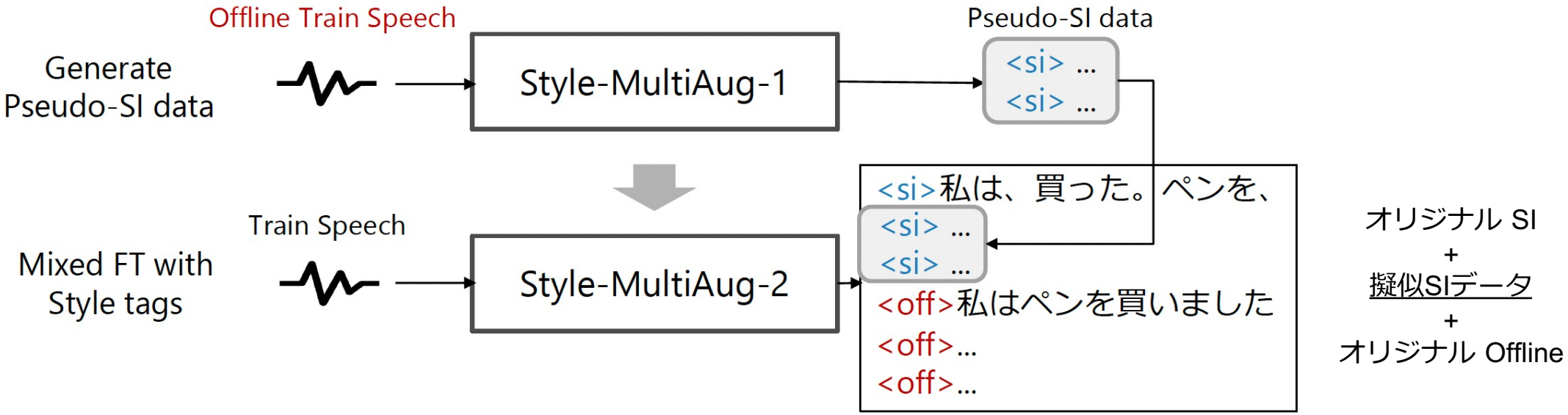
- 学習時
  - スタイルタグを目的テキストの先頭につける
- 推論時
  - 出力したいスタイルのタグを与えて、Forced Decodingによりそのスタイルの出力を得る
- 目的
  - 少量のSIデータだけでなくOfflineデータも用いつつ、スタイルの区別ができるモデル (SIデータのデータ少量問題の軽減)



# 提案法+
# スタイルタグ付きデータでの混合学習+自己学習 (Style Multi-Aug FT)

■ 1度提案法で学習したモデルから擬似SIデータを生成

→ そのデータで提案モデルを再度学習 （これを収束するまで繰り返す）



# 実験設定

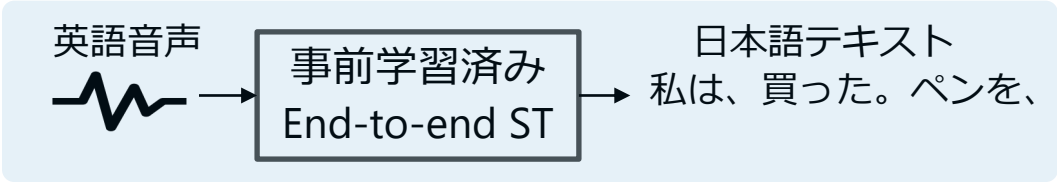
- 従来法：それぞれのデータで事前学習済みモデルを直接Fine-tuning
  - **Offline FT**
  - **SI FT**
  - **Mixed FT**

- 提案法 : スタイルタグ付きデータでの混合学習 (Style FT)
  - **Style FT**

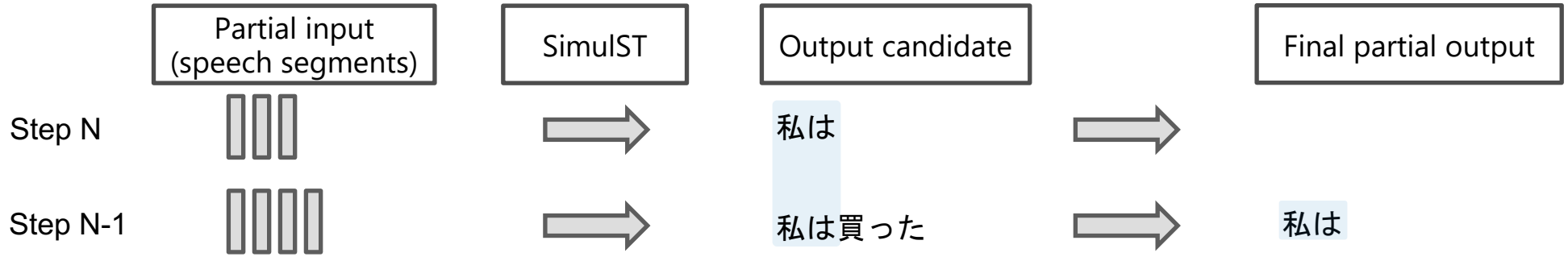
- 提案法+ : スタイルタグ付きデータでの混合学習+自己学習 (Style Multi-Aug FT)
  - **Style-MultiAug-4 FT** (開発データで損失最小のN=4回目を選択)

# 実験設定

- データ
  - **Offline**: MuST-C TED En-Ja [Di gangi+2019] (Train: 328.6k, Test: 2841)
  - **SI**: NAIST-SIC-Aligned-ST En-Ja [Ko+2023] (Train: 65k, Test: 511)
  - **CMT test** set (511) [Fukuda+2024]
- 事前学習済みEnd-to-end ST: HuBERT Encoder + mBART Decoder




- 部分出力のための訳出方法：Local Agreement [Liu+2020]
  - 現時点の出力候補と前時刻出力の共通部分を出力として確定



# 実験設定

## ▷性能評価 (Y)

**BLEU** 日本語出力 ↔ 日本語の正解文

**BLEURT** 日本語出力 ↔ 日本語の正解文

**COMET-QE** 日本語出力 ↔ 英語の入力ソース文

**表層一致度計算**

**文のEmbeddingベースで類似度計算**

## ▷遅延評価 (X)

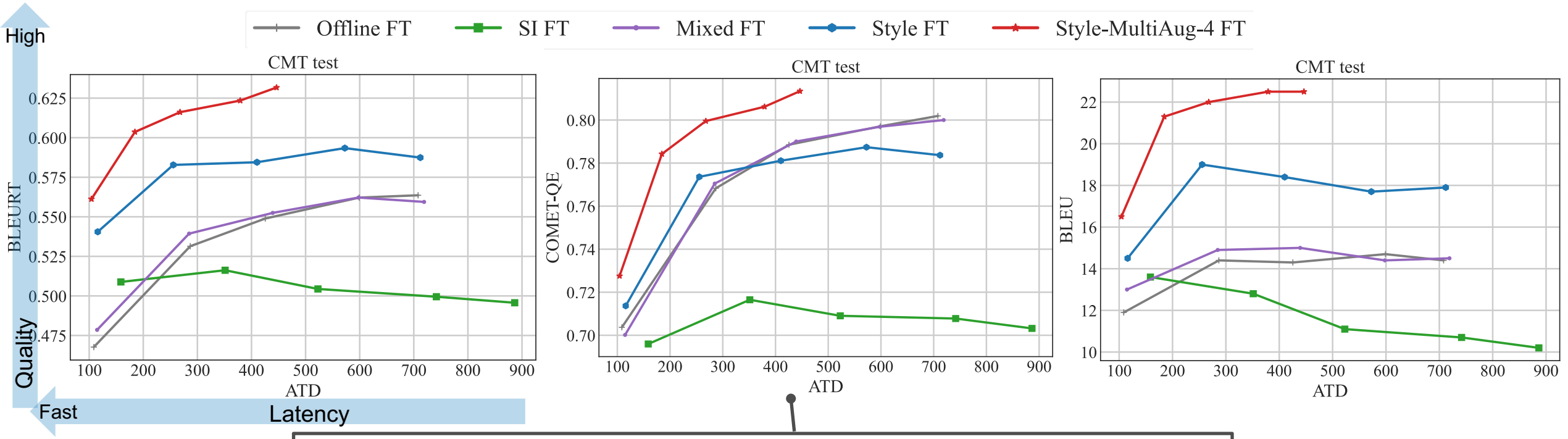
**Average Token Delay (ATD)** [Kano 2023+] 訳出終了時刻を考慮した遅延指標 (単位:ms)

# 評価結果

■ 両提案法 Style-MultiAug-4 FT  Style FT で品質と遅延が向上

- 提案法+が提案法より良い : 擬似データでの多段階Tuningの効果が見られる

Test:　CMT test
Metric: BLEURT, COMET-QE, BLEU




- COMET-QEの向上幅が大きい (ソース入力テキスト↔出力の評価)
→ 提案法は入力の内容を保持できる

# 出力文の例

■ 提案法 Style FT, 提案法+ Style-MultiAug-4 FT: 冗長な部分があり、出力が長い傾向

- 従来法 SI FTより話し手の内容を保持した出力が可能

通訳者による人手評価スコア (1-5)

| **Example 1** | | **Adequacy** | **Fluency** |
|---|---|---|---|
| Source | It's probably the smallest of the 21 apps that the fellows wrote last year. | | |
| SI FT (従来法) | 一番小さいアプリです。 | 3 | 3 |
| Style FT (提案法) | 恐らく21のアプリの中で、一番小さいものだと思います。 | 4 | 4 |
| Style-MultiAug-4 FT (提案法+) | これは、おそらく、21のアプリの中で、最も小さいものです。昨年、フェローが書いたものです。 | 5 | 4 |
| SI test reference | 昨年作ってくれたもので。 | | |
| CMT test reference | おそらくそれは最小です、21のアプリの中で、昨年フェローが書いたものの中で。 | | |
| **Example 2** | | | |
| 入力音声の話している内容 | It was running into bankruptcy last fall, because they were hacked into. | | |
| SI FT (従来法) | 破産したんです。この秋に破産したんです。 | 3 | 4 |
| Style FT (提案法) | これは、去年の秋に、破産したものです。なぜなら、彼らは、ハッキングされたからです。 | 5 | 5 |
| Style-MultiAug-4 FT (提案法+) | それは、昨年、破産につながったものです。なぜなら、彼らは、不正に侵入されたからです。 | 5 | 5 |
| SI test reference | 破産をしたのは、去年の秋なんです。ハッキングをされたからです、 | | |
| CMT test reference | それは、昨秋、破産寸前でした、ハッキングされたためです。 | | |

従来法では必要な情報が欠落
極端に短い出力

「それは」「なぜなら」「彼らは」
→ 冗長性が高く、省略されても問題ない

# まとめと今後の方向性

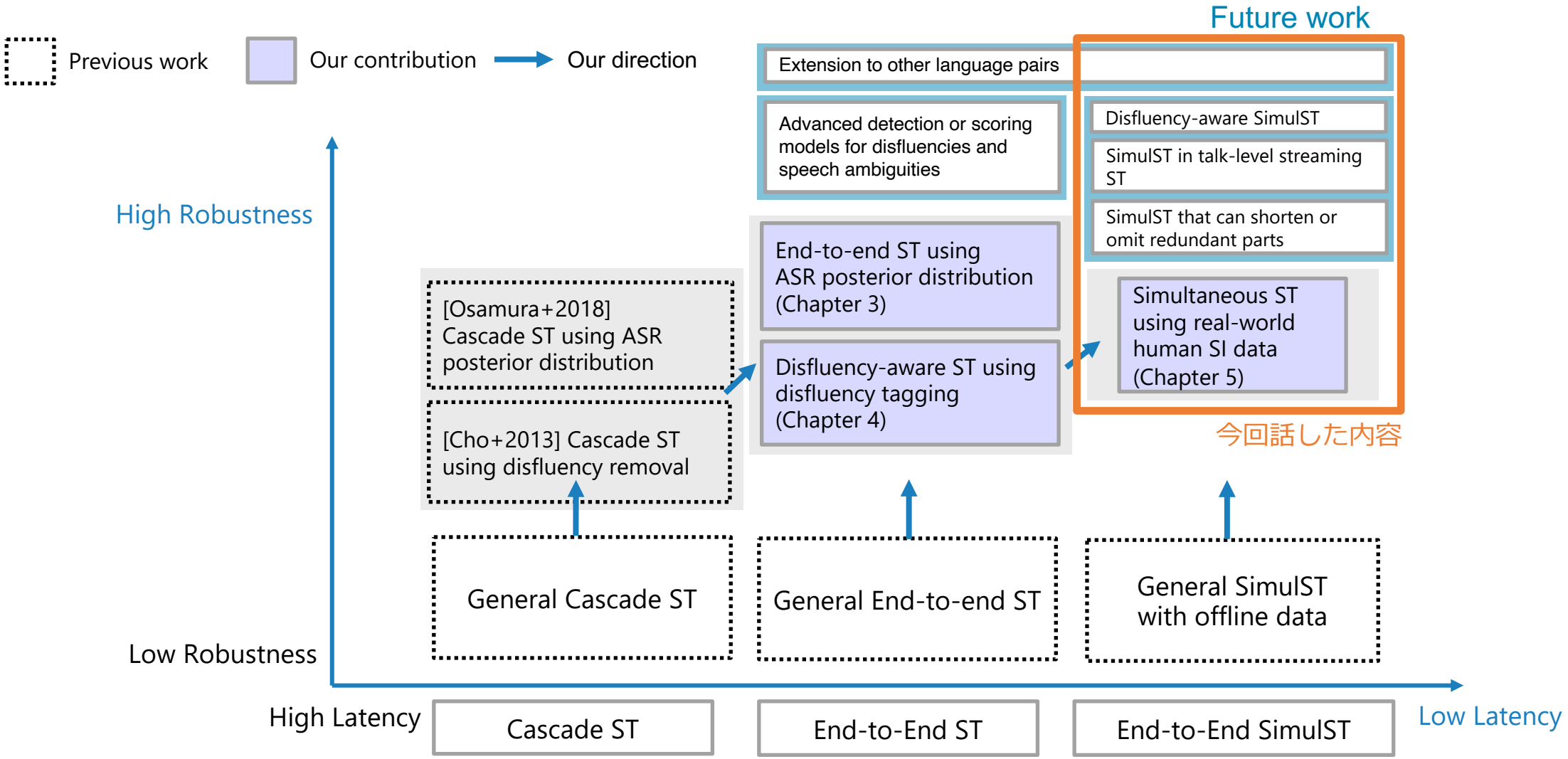
■ 背景
- 近年の同時音声翻訳は翻訳データを用いておりタスクとのミスマッチがあった

■ 提案法
- 同時通訳データを用いた同時音声翻訳モデル
- データ少量問題を軽減する効果的な学習手法の提案

■ 結果
- 提案法は話し手の話す内容を保持しつつ、低遅延かつ高性能な翻訳が可能

■ **今後の方向性**
- トーク単位で、話し手の発話に追いつけるかの評価 (今回：文単位で独立した評価)
- 同時通訳者のように適切な省略が可能なモデル（今回：提案法で冗長な部分が見られる）
- 曖昧な入力に頑健な同時音声翻訳モデル

# 今後とロードマップ


Previous work
Our contribution
Our direction
Future work
Extension to other language pairs
Advanced detection or scoring models for disfluencies and speech ambiguities
Disfluency-aware SimulST
SimulST in talk-level streaming ST
SimulST that can shorten or omit redundant parts
End-to-end ST using ASR posterior distribution (Chapter 3)
Simultaneous ST using real-world human SI data (Chapter 5)
[Osamura+2018] Cascade ST using ASR posterior distribution
Disfluency-aware ST using disfluency tagging (Chapter 4)
[Cho+2013] Cascade ST using disfluency removal
今回話した内容
High Robustness
Low Robustness
General Cascade ST
General End-to-end ST
General SimulST with offline data
High Latency
Low Latency
Cascade ST
End-to-End ST
End-to-End SimulST

# Thank you

# Publications

End-to-end Simultaneous Speech Translation with Style Tags using Human Simultaneous Interpretation Data, Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, Sakriani Sakti and Satoshi Nakamura Journal of Natural Language Processing, Vol.32, No.2, To appear. Jun. 2025.

Neural End-to-end Speech Translation Leveraged by ASR Posterior Distribution Yuka Ko, Katsuhito Sudoh, Sakriani Sakti and Satoshi Nakamura IEICE TRANSACTIONS on Information and Systems, Vol.E107-D, No.10, pp. 1322-1331. Oct. 2024.

Word Order in English-Japanese Simultaneous Interpretation: Analyses and Evaluation using Chunk-wise Monotonic Translation. Kosuke Doi, Yuka Ko, Makinae Mana, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 21th International Conference on Spoken Language Translation (IWSLT), pp. 254-264, Bangkok, Thailand, Aug. 2024.

NAIST Simultaneous Speech Translation System for IWSLT 2024 Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Haotian Tan, Makoto Sakai, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 21th International Conference on Spoken Language Translation (IWSLT), pp. 170–182. Bangkok, Thailand, Aug. 2024.

NAIST-SIC-Aligned: Automatically-Aligned English-Japanese Simultaneous Interpretation Corpus. Jinming Zhao*, Yuka Ko*, Kosuke Doi, Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pp. 12046- 12052, Torino, Italia, May 2024. (*: Equal Contribution)

Tagged End-to-End Simultaneous Speech Translation Training using Simultaneous Interpretation Data Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT), pp. 363–375. Toronto, Canada, Jul. 2023. (Best Student Paper Award)

NAIST Simultaneous Speech Translation System for IWSLT 2023 Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT), pp. 330–340. Toronto, Canada, Jul. 2023.

NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022 Ryo Fukuda, Yuka Ko, Yausumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT), pp. 286–292. Dublin, Ireland, May. 2022.

# Appendix

# Selected Reference

[Fukuda+2022] Ryo Fukuda, Yuka Ko, Yasumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2022. NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 286–292, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

[Osamura+2018] Kaho Osamura, Takatomo Kano, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Using Spoken Word Posterior Features in Neural Machine Translation. Proceedings of the 15th International Workshop on Spoken Language Translation, 181-188, Oct. 2018.

[Weiss+2017] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. In Francisco Lacerda, editor, Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, pages 2625–2629. ISCA, 2017.

[Inaguma+2020] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and ShinjiWatanabe. ESPnet-ST: All-in-One Speech Translation Toolkit. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, pages 302–311. Association for Computational Linguistics, 2020.

[Cho+2023] Eunah Cho, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. A corpus of spontaneous speech in lectures: The KIT lecture corpus for spoken language processing and translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1554– 1559, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[Di gangi+2019] Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

# Selected Reference

[Tsiama+2022] Ioannis Tsiamas, Gerard I. Gállego, Carlos Escolano, José Fonollosa, and Marta R. Costa-jussà. Pretrained speech encoders and efficient fine-tuning methods for speech translation: UPC at IWSLT 2022. In Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà, editors, Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 265–276, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[Doi+2021] Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In Marcello Federico, Alex Waibel, Marta R. Costa-jussà, Jan Niehues, Sebastian Stuker, and Elizabeth Salesky, editors, Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), pages 226–235, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics.

[Ko+2023] Yuka Ko, Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 363–375, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics.

[Liu+2020] Liu, Danni, Gerasimos Spanakis, and Jan Niehues. "Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection." (2020).

[Kano+2024] Yasumasa Kano, Katsuhito Sudoh, and Satoshi Nakamura. Average Token Delay: A Duration-aware Latency Metric for Simultaneous Translation. Journal of Natural Language Processing, 31(3), 2024.

[Fukuda+2024] Ryo Fukuda, Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 原発話に忠実な英日同時機械翻訳の実現に向けた順送り訳評価データ作成 [creation of evaluation data for monotonic translation toward the realization of simultaneous english-japanese machine translation faithful to the source speech]. In Proceedings of the 259th meeting of Special Interest Group of Natural Language Processing (IPSJ-SIGNL), 2024-NL-259(14), pages 1–6, 2024. (in Japanese).

# 本発表での補足資料

# 評価データ

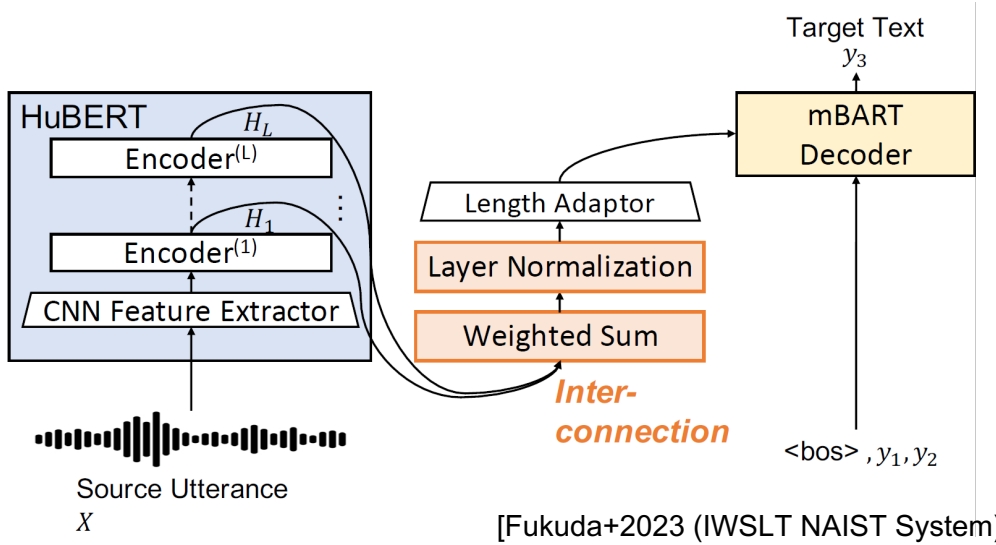
- SI test (511):
  - NAIST-SIC-Aligned-ST test
- Offline test (2841)
  - MuST-C TED tst-COMMON
- CMT test set (511) [Fukuda+2024]
  - 順送り訳テストデータ Chunk-wise Monotonic Translation data
    - Target to make “high quality chunk-wise monotonic translation data (順送り訳)” without omission effect

**Source:** To show you now / what we are working on / by starting out talking / about the American solder, / that on average / does carry out 100 lbs.
**Target:** これからお伝えする上で，/ 私たちの取り組みをお伝えする上で，/ 初めに話すのは，/ アメリカの兵士についてです，/ 平均して / 彼らは約100ポンドの荷物を運びます.

| | Naturalness | Word order difference (serious for latency) | Omission |
|---|---|---|---|
| A: Offline ST (MuST-C) | ◎ | Large | Few |
| B: true SI data (NAIST-SIC-Aligned-ST) | △ | Small - Large | Many |
| C: chunk-wised monotonic data by MT | × | Small | Few |
| D: chunk-wised monotonic data by human | ◯ | Small | Few |

# Experiment setting

- Data
  - Offline: MuST-C En-Ja [Di gangi+2019]
  - SI: NAIST-SIC-Aligned-ST En-Ja [Ko+2023]
- Pretrained offline ST model
  - HuBERT+mBART model [Fukuda+2023]
- Simultaneous decoding
  - Local Agreement [Liu+2020]
  - Speech segment size: {200, 400, 600, 800, 1000}ms
  - Style tag in inference step
    - SI Test: output from <si> tag
    - Offline test: output from<off> tag
- Evaluation tool
  - SimulEval




[Fukuda+2023 (IWSLT NAIST System)]

BLEURT
- The **sentence semantic similarity** between hypothesis and reference

ATD (Average Token Delay)
- Latency metric **focuses on the end timings of partial translations**

# (1) 話し言葉に含まれる曖昧な入力に強いEnd-to-end音声翻訳

# 背景：曖昧な入力による音声翻訳誤り

■ とある出力例 (En-Ja End-to-end offline ST in IWSLT2022) [Fukuda+2022]

◇モデルの目的言語出力 (日本語)
これがその夜の真っ只中でした **(1)プール**の外で **(2)ナブララ**とパジャマの裸足で 炎に包まれていました

◇DeepL (英語)
This was in the middle of that night, outside the **(1) pool**, with **(2) Navrara**, barefoot in her pajamas, engulfed in flames.

- **(1)「pouring** rain」→「**プールの外に」: pouring** is similar **pool**?
- **(2)「under an umbrella」→「ナブララ: "navrara"」**: by hesitation?

---

◇原言語文 (英語)
Here it was, the middle of the night, she was standing outside in the **(1) pouring rain**, **(2) under an umbrella**, in her pajamas, barefoot, while her house was in flames.

◇参照訳 (日本語)
真夜中でしたが 家主の女性は 炎に包まれている家の前で **(1)打ちつける雨の中 (2)傘を差し** パジャマを着て裸足で立っていました

→ 曖昧性に頑健なEnd-to-end音声翻訳モデルの提案

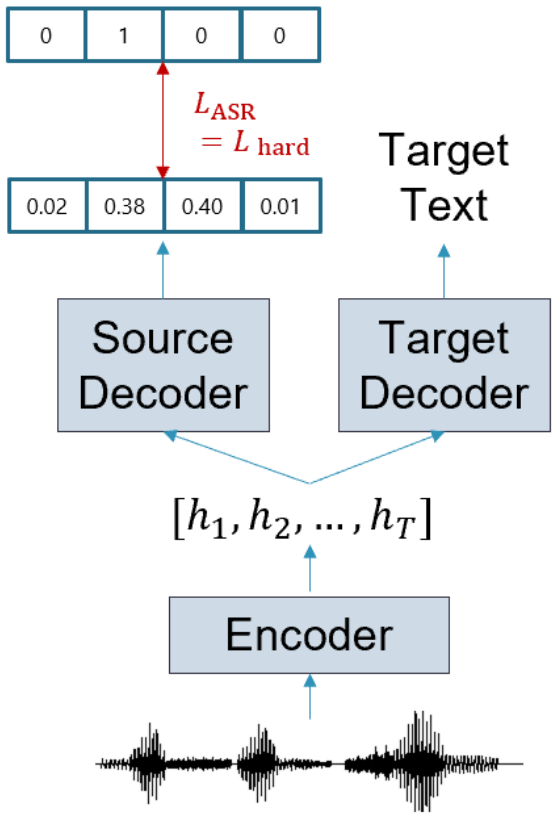
# 従来モデルと提案モデル

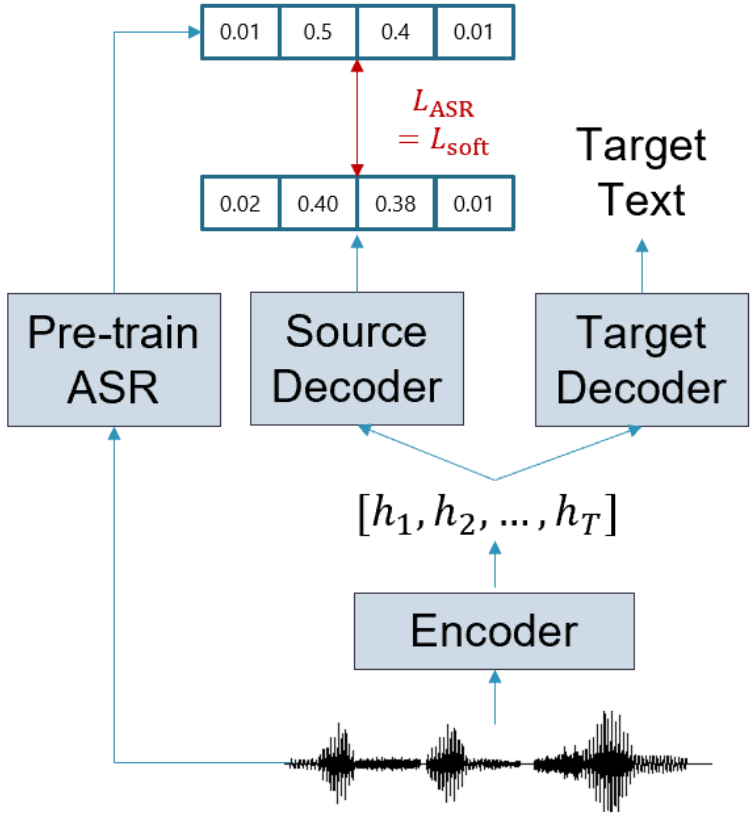
- 従来モデル関数：$L_{\text{hard}}$
- 学習時の参照訳：One-hotの参照訳

△ 誤りに対して同じスコアが与えられる
△ 発音の曖昧性を考慮できない

- 提案モデル関数：$L_{\text{soft}}$ [Ko+2021]
- 学習時の参照訳：**音声認識(ASR)の出力確率分布**

〇発音の曖昧性（聞き間違えやすさ）を考慮可能




答えの参照訳
I catch a ball
予測パターン1
I cat a ball
予測パターン2
I lost a ball

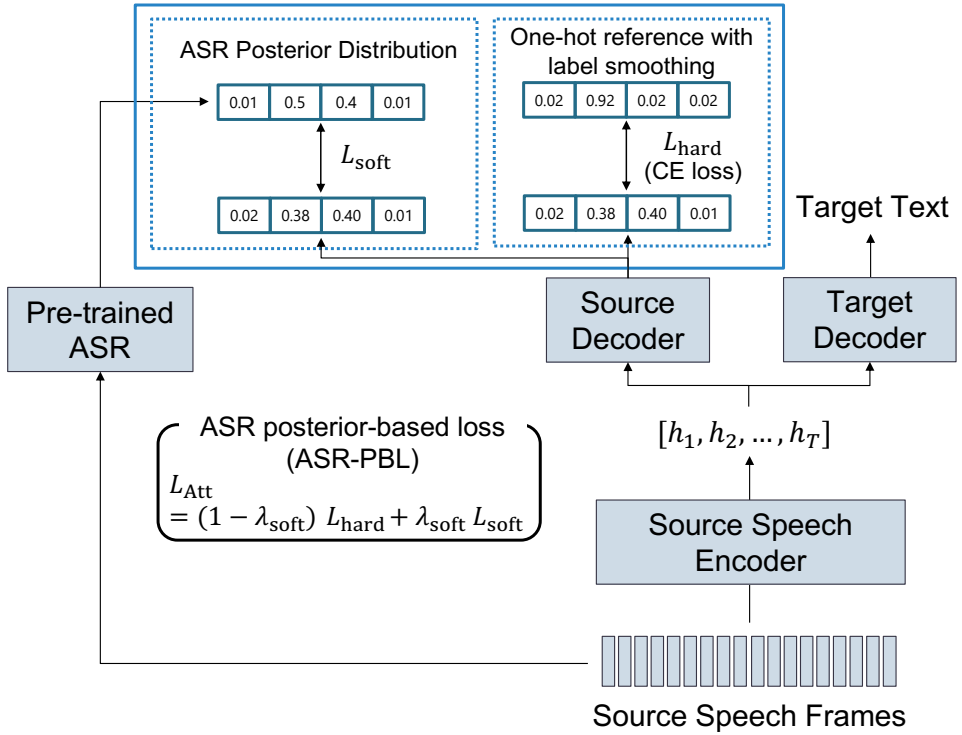
# Experiment setting

- Data: Fisher Spanish Corpus
  - Spanish Speech→ English Text (include disfluencies)
- ST model
  - Transformer-based models in ESPnet
  - ST-task loss $L_{\mathrm{ST}}$ : CE
  - **ASR-task loss $L_{\mathrm{ASR}}$**
    - Baseline : $L_{\mathrm{ASR}}$ : CTC+ CE ($\boldsymbol{\lambda_{ctc}} = \mathbf{1.0, 0.5, 0.3}$)
      - Based on Hybrid/CTC Attention loss
    - **Proposed : ASR Posterior-based Loss (ASR-PBL)**

$$\boldsymbol{\lambda_{\mathrm{ASR}}} = \{\mathbf{0.3, 0.4, 0.5}\}, \boldsymbol{\lambda_{\mathrm{soft}}} = \{\mathbf{0.1, 0.3, 0.5, 0.7, 0.9, 1.0}\}$$

$$\boldsymbol{L_{\mathrm{Att}} = \lambda_{\mathrm{soft}} L_{\mathrm{soft}} + (1 - \lambda_{\mathrm{soft}}) L_{\mathrm{hard}}}$$

- $\boldsymbol{L = \lambda_{\mathrm{CTC}} L_{\mathrm{CTC}} + (1 - \lambda_{\mathrm{CTC}}) L_{\mathrm{Att}}}$
- $\boldsymbol{L = \lambda_{\mathrm{ASR}} L_{\mathrm{ASR}} + (1 - \lambda_{\mathrm{ASR}}) L_{\mathrm{ST}}}$



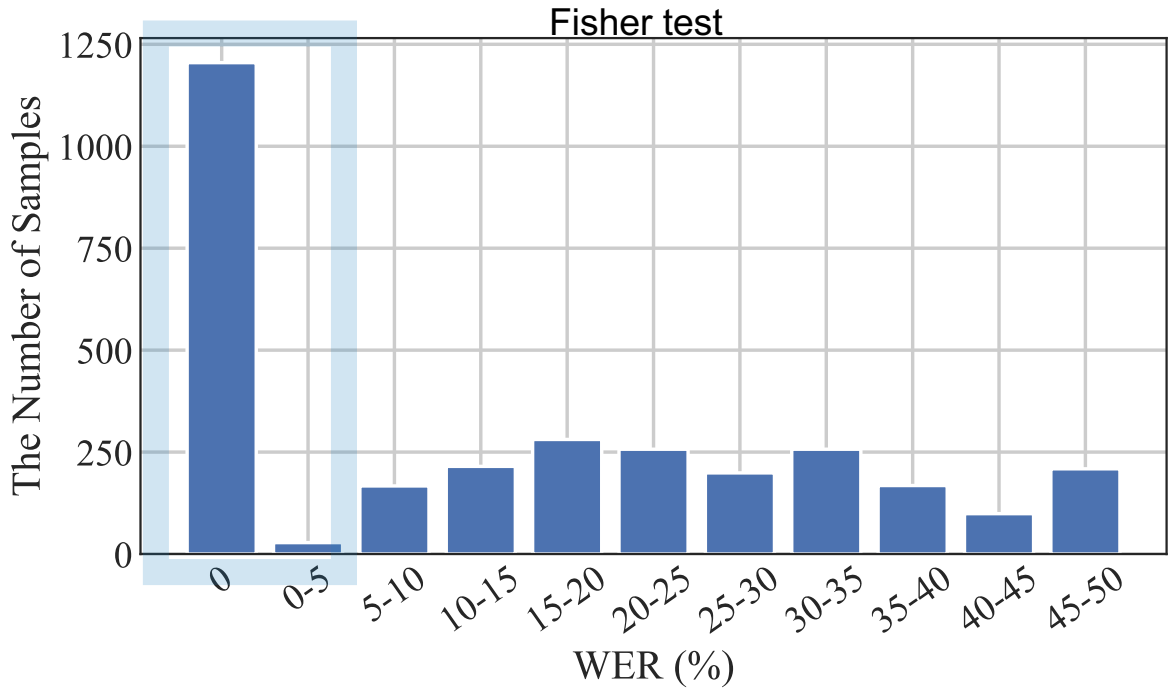
# Results

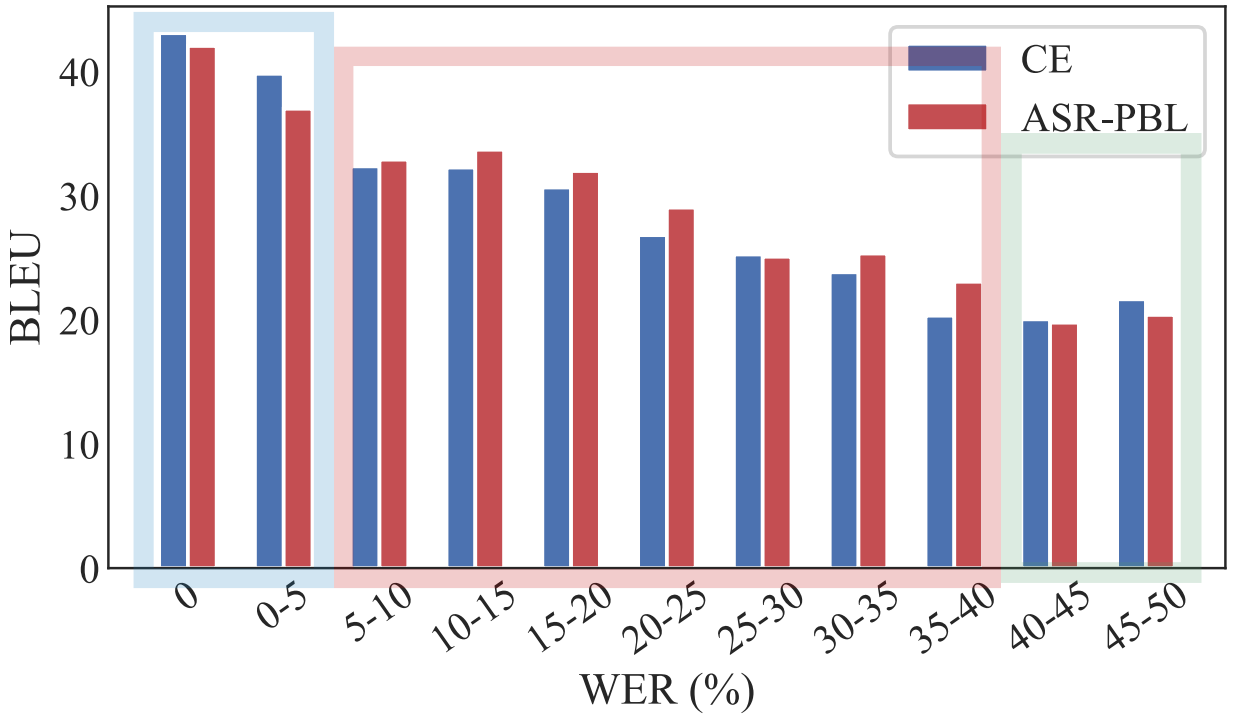
- Our proposal surpassed other baselines

| **Model** | | | | | | **BLEU** | |
|---|---|---|---|---|---|---|---|
| Task | ASR task loss | $\lambda_{CTC}$ | $\lambda_{ASR}$ | $\lambda_{soft}$ | Dev | Dev2 | Test |
| Single-task ST | - | - | - | - | 41.10 | 41.61 | 40.66 |
| Multi-task ST | ESPnet Transformer ASR-MTL [Inaguma+2020] | 0.3 | 0.3 | - | 46.64 | 47.64 | 46.45 |
| | w/o CE ($\lambda_{CTC} = 1.0$) | 1.0 | 0.3 | - | 45.98 | 47.16 | 45.83 |
| | CE ($\lambda_{CTC} = 0.5$) | 0.5 | 0.5 | - | 47.18 | 47.43 | 46.59 |
| | ASR-PBL ($\lambda_{CTC} = 0.5$) | 0.5 | 0.3 | 0.7 | **47.20** | **48.36** | **46.82** |

# Discussion: BLEU Comparison across various WER ranges

- Motivation: Our ST can treat high-ambiguous input?
  - ASR difficulty is high (WER is high) → input speech will be more ambiguous (difficult to be translated)
- 5-40% : proposed ST achieved higher BLEU scores
- Low WER 0-5%, high WER (40%-): no improvement
  - but sample count 0% - 5% range was very small & these samples contained minimal ASR errors



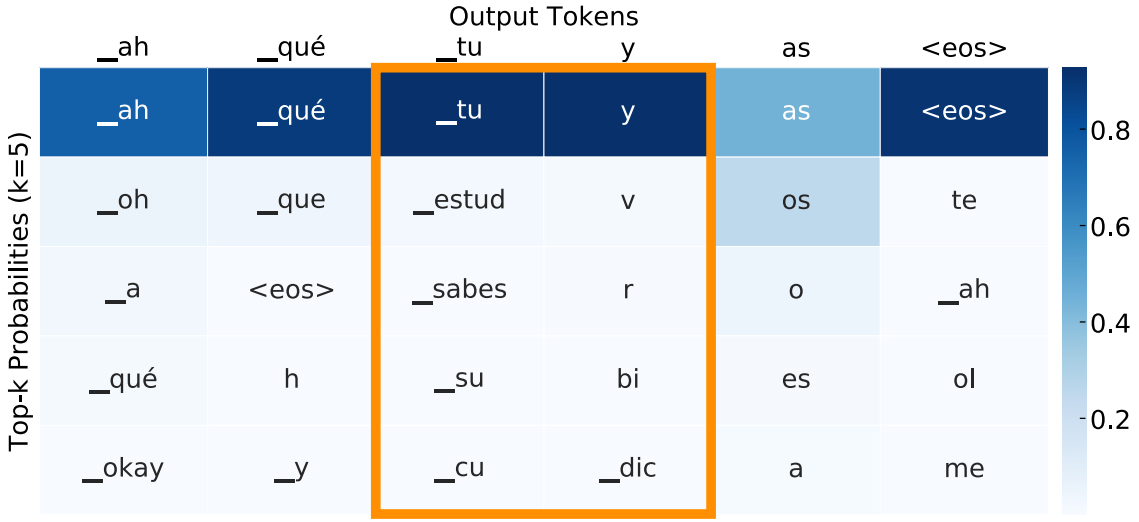



# Output example

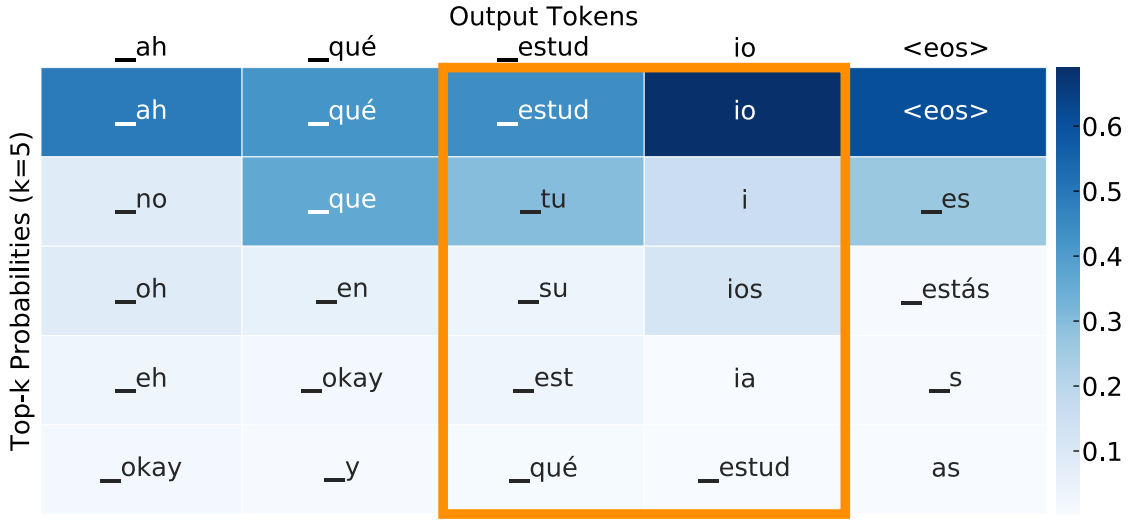
- Proposal gives scores not only to 1-best **estudio**, but also to other candidates *tu, i, ios*

| **Example** | ASR task output | ST task output |
|---|---|---|
| Reference | ah qué *estudias* | oh what do you *study* |
| CE (Baseline) | ah qué tuyas | ah what are you doing |
| ASR-PBL (Proposal) | ah qué **estudio** | oh what do you **study** |

Top-5 ASR task posterior probabilities (first raw: 1-best output tokens)




CE ASR task (Baseline)




ASR-PBL ASR task (Proposal)