訳出の同時性に特化した評価データを用いた同時音声翻訳モデルの評価と分析

蒔苗 茉那　坂井 優介　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{makinae.mana.mh2,sakai.yusuke.sr9,kamigaito.h,taro}@is.naist.jp

概要

同時音声翻訳は，原言語文のセグメント化や目的言語文を原言語文の語順に近づけることで，低遅延と高品質の両立を目指してきた．しかし，既存の評価データは語順の並び替えを多く含むため，低遅延の同時音声翻訳評価に適していない．本研究では，目的言語文が原言語文の単語・句の並びになるべく沿うような語順の単調性に焦点を当てた新しい評価データを提案する．実験の結果から，提案評価データ *simul-tst-COMMON* は，既存の評価データよりも適切にモデルの性能評価ができることを示した．

1　はじめに

同時通訳 (SI) にて，同時通訳者 (通訳者) は原発話を小さな単位に分割し逐次的に処理する (サラミ技法 [1]) ことで，目的発話の語順変更を最小限に抑え，原言語に対する目的言語側の語順の単調性（単調性）を維持しながら低遅延かつ高品質な訳出を行なっているとされる [2]．同時音声翻訳 (SiST) や同時翻訳 (SiMT) でも同様に, 品質と遅延のバランスをとるために，原言語文の文分割の最適化 [3, 4, 5] や，原言語文の語順にできるだけ沿うよう目的言語文を書き換える方法が提案されている [6, 7, 8]．

一方で，現在 SiST や SiMT の評価時に使用されるテストデータは，一文の入力を終えてから訳出を開始する音声翻訳評価で使用されるテストデータと同じであることから，訳出の同時性が求められる SiST や SiMT の評価に適さない可能性が指摘されている [9, 8]．また，実際の SI の書き起こしをテストデータとして用いた研究があるが [10, 11]，SI には訳出の誤り等が含まれているため，評価の適切性に欠くことが指摘されている [9, 12]．

そこで，SiMT の評価に特化したテストセットが提案され [13], その効果も検証されている [7, 9]．このテストデータは，発話をより短い単位であるチャンクに区切りながら順次訳出する順送り方略 [14] をルールベースで定義し，そのルールに基づいて英→日で作成している．しかし，このテストデータは言語ペアに限りがあることやテストデータの大きさ，そして理想的な単調性の程度が明らかでない．

本研究では，SiST や SiMT に特化した評価データ *simul-tst-COMMON* を提案する．本テストセットは，既存の音声翻訳の評価データ tst-COMMON の英→{ 日, 中, 独 } を対象に，単語・句の単調性に焦点を当て，音声の書き起こし，原言語の語順になるべく沿った翻訳，通訳者による品質チェックを通じて作成している．実験結果から，既存の評価データよりも *simul-tst-COMMON* の方が SiST 評価に適していることを示した．また分析の結果から，既存の品質・遅延評価の限界も明らかになった．

2　関連研究

一文全体の音声入力を終えてから翻訳を開始する音声翻訳と異なり，SiST や SiMT は一文を部分的かつ逐次的に翻訳するため，低遅延と高品質の両立が求められる [15, 16]．そこで主に 2 つのアプローチが提案されている．ひとつは，最適な文分割の場所を決定するために出力のタイミングを改善する手法で，もうひとつは訳出の同時性に対応するために学習データの目的言語の書き換えを行う手法である．

出力のタイミングの改善を目指す初期の研究では，韻律 [3] や語彙情報など [4, 5] を元に，一文をより細かく分割する場所を決定している．近年では，特定の単語数や文字数を読み込んだ後に入力と出力を交互に行う固定型 [16] や，文脈に基づいて最適な分割場所を決定する適応型 [17] が提案されている．また，原言語文の構文や特定の出力のタイミング手法に合わせるために，学習データの目的言語を擬似的に書き換える手法も存在する．これには，文法的

および意味を保ちながら構文変換ルールを定義し適用したり [6]，大規模言語モデル (LLM) を活用して通訳者が行うような訳出スタイルを模倣するプロンプトを与えたりする手法 [7, 8] がある．

そのほか，訳出の同時性が求められるタスクに適した評価データの必要性も指摘されている．先行研究には，SiMT は実際の SI データで評価すべきいう考え [11] に基づき，SI 書き起こしをテストデータとして用いた研究がある [7, 9, 10, 18]．しかし，通訳者の訳出には誤り等が含まれ [12]，それらが原因でモデルの性能が過小評価される可能性も指摘されている [9]．そこで，順送り方略 [14] のみを適用したテストセットが提案 [13]・検証されている [7, 9] が，対象が英日のみであるなど課題が残されている．

3 評価データについて

構築の流れ 本研究では，英→ { 日, 中, 独 } に注目し，MuST-C v2.0 [19] の tst-COMMON を元に評価データを構築した．これら言語ペアは，原言語に対する目的言語の語順差の程度が異なることや，元にしたデータが IWSLT2024 Simultaneous track で使用されている [20] ことから選択した．MuST-C v2.0 は，英語の講演音声とそれに対応付いた書き起こしテキスト，各言語の翻訳テキストからなる．

第一に，Whisper の medium.en モデル [21] を用いて講演音声の書き起こしを行った．これは，元の書き起こしテキストに含まれる *Laughter* や *Applause* 等の非言語要素を取り除き，発話内容のみの訳出を目指すからである．書き起こしテキストを人手で確認した後，ピリオドやクエスチョンマークを元に一文単位で分割した．最後に，Gentle [1] を使って音声とのアライメントを取った．その結果，各言語ペアで行数が 2,806 に統一された．第二に，サラミ技法を元に, 原言語文をより小さな単位に分割して逐次処理するよう，LLM[2] にプロンプトを与えて翻訳を行った [8]．このように，SI の訳出技法や LLM を用いることで，単調性の確保やスタイルの統一，訳抜けなしの翻訳が可能となるので，既存の自動評価指標との整合性が高まることが期待される．第三に，通訳者が自身の経験に基づき，第二ステップで生成された翻訳を，品質・単調性・適切性・用語の一貫性の 4 つの観点から品質を確認し，品質が低い場合には修正を行なった．

1） https://github.com/lowerquality/gentle
2） GPT-4o [22] を使用した．

表 1 語順の単調性の比較

言語ペア	simul-LLM	simul-tst-COMMON	offline-LLM	tst-COMMON
英日	0.795	**0.720**	0.587	**0.522**
英中	0.948	**0.896**	0.874	**0.875**
英独	0.969	**0.962**	0.943	**0.938**

表 2 通訳者による修正前後の訳出の品質比較

言語ペア	BLEU	ChrF	TER
英日	82.6	85.3	62.2
英中	78.3	73.4	68.6
英独	96.5	98.0	2.97

データ分析 構築した評価データについて，単調性の程度・品質・事例の分析を行った．比較対象は，通訳者による修正前の simul-LLM，修正後の simul-tst-COMMON，一般的な翻訳の指示を LLM に与えて生成した offline-LLM である．他の 3 つと行数が異なるものの，比較のため tst-COMMON でも単調性の程度を計算した．

単調性の程度は，Awesome Alignment [23] から取得した原言語文と目的言語文間の単語のアライメントを元に，スピアマンの順位相関係数を使って求めている．単調性が 1 に近いほど語順変更が少ないことを表し，遅延と品質のバランスを取るのに有益であることを意味する．表 1 に示すように，simul-LLM は最も高い単調性を達成している．通訳者による修正が行われた simul-tst-COMMON は，simul-LLM と比較して単調性が低下しているが，これら単調性の低下は，通訳者による修正を経て品質の向上に寄与しているので許容されるべきである．simul-tst-COMMON と tst-COMMON・offline-LLM を比較すると, 特に英日にて単調性の差が顕著である．これは，文法的な構造の差が増すにつれて，SiST や SiMT の難しさが増すことを示唆する．また，SiST や SiMT において，従来使われてきた tst-COMMON は適切な評価が難しい可能性を示唆する．

表 2 は，通訳者による修正前後の simul-tst-COMMON の品質比較である．構築プロセスにて，元のテストデータと比べて行数が変化したため，tst-COMMON との直接的な比較ではなく，修正前を参照訳，修正後を生成文とし，BLEU [24]，chrF [25]，TER [26] で評価した．BLEU と chrF から，修正前後で文の意味が保持されていることが確認できる．また TER から，英日・英中において通訳者が自身の経

表 3 英日における SI およびオフライン設定での句の順序差の比較例.

原言語文	(1) But still it was a real footrace / (2) against the other volunteers / (3) to get to the captain in charge / (4) to find out what our assignments would be.
simul-LLM	(1) それでも，本当に競争でした (*But still it was a real footrace*). / (2) 他のボランティアと (*against the other volunteers*) / (3) 担当のキャプテンに会うために (*to get to the captain in charge*), / (4) 私たちの任務が何であるかを知るために (*to find out what our assignments would be*).
simul-tst-COMMON	(2) それでも，まさに，他のボランティアとの (*against the other volunteers*) / (1) 競争でした (*it was a footrace*). / (3) 担当の消防隊長に会って (*to get to the captain in charge*), / (4) 自分の任務が何か確かめるのは (*to find out what our assignments would be*).
offline-LLM	(4) それでも，私たちの任務が何であるかを知るために (*to find out what our assignments would be*), / (3) 責任者であるキャプテンのところに行くのは (*to get to the captain in charge*), / (2) 他のボランティアとの (*against the other volunteers*) / (1) 本当の競争でした (*it was a real footrace*).
tst-COMMON	(3) それでも団長を見つけて (*to get to the captain in charge*) / (4) 任務を割り振ってもらうのに (*to find out what our assignments would be*) / (2) 他のボランティアと (*against the other volunteers*) / (1) 激しい競走になりました (*it was a real footrace*).

験を基に，実際の通訳のシナリオに近づけるために生成された翻訳を頻繁に修正していることが示されている．一方，英独で TRE が小さいのは，英語と文法的に距離が近いことが原因と考えられる．

表 3 は英日間の SI(simul-LLM・simul-tst-COMMON) およびオフライン (offline-LLM・tst-COMMON) 設定での句の順序差の程度を比較したものである．ここでは，LLM による単語や句の順序を単調に調整した結果が，通訳者の視点から見て自然であるかを分析することを目的とした．原言語の末尾にある句 (4) が，tst-COMMON や offline-LLM では目的言語の冒頭に配置されている．これは，SI において，このような長距離の語順変更が適切でないことを示す一例である．一方，simul-LLM や simul-tst-COMMON では，原言語の句の位置により忠実に訳出されており，語順変更が最小限に抑えられていることが確認された．このような訳は，目的言語側で流暢性を損なう可能性があるものの，訳出の同時性が求められるタスクでは不可欠である．また，通訳者は，LLM による単語や句の単調性を常にそのまま受け入れるわけではなく，単調性が過度である場合は適宜修正を加えている．句 (1) にあるように，simul-LLM でみられた LLM による過度な語順の単調性に対し，通訳者の判断を反映して修正が行なわれたことで，評価データとしての品質が維持されている．

4 実験設定

実験では，異なる特徴を持つ学習データの違いが，既存の tst-COMMON と提案する simul-tst-COMMON の間で評価結果にどのような差をもたらすのかを検証する．具体的には，頻繁な語順の並び替えを含むオフライン翻訳スタイルの MuST-C [19] と，目的言語を原言語の語順に近づけた訳出スタイルの Simul-MuST-C [8] で学習させたモデル間の差を比較する．NAIST IWSLT 2023 system paper [27] を参考に，fairseq [28] を使用し，モデルアーキテクチャは Transformer [29] を用いた．エンコーダの初期化は事前学習済み音声モデル HuBERT-Large [30] の，デコーダは多言語モデル mBART50 [31] の重みを使用した．デコーディングは wait-k [16] を使い，k=1 で読み込む固定長セグメントを 160ms とし，k の値を {3,5,7,9,11,13,15,17} に設定した．品質は，BLEU [24]，BERTScore [32]，BLEURT [33]，COMET [34] で，遅延は，Average Lagging(AL) [16]，Length Adaptive Average Lagging(LAAL) [35]，Average Token Delay(ATD) [36] で評価した．

5 実験結果

図 1 にあるように，英日の場合，simul-tst-COMMON では，オフライン翻訳スタイルである MuST-C と，目的言語の語順を原言語の語順により近づけた訳出スタイルである Simul-MuST-C をそれぞれで学習させたモデル間で，顕著な品質差が確認できる．tst-COMMON でも学習データの特徴の違いが結果に現れているが，simul-tst-COMMON の場合と比べて，その差は小さい．さらに，simul-tst-COMMON で評価した際のスコア範囲とは異なり，tst-COMMON ではすべてのモデルで性能が低い結果となっている．遅延でも，Simul-MuST-C が優れていることを示す．これら異なるテストセット間で見られる品質の差から，同時性が求められるタスク向けのモデルは，目的言語の語順を原言語の語順により近づけた訳出スタイルのように，同時性を達成するために必要な特徴を反映したテストデータで評価する必要があると考えられる．そのため，simul-tst-COMMON は，tst-COMMON よりも同時性

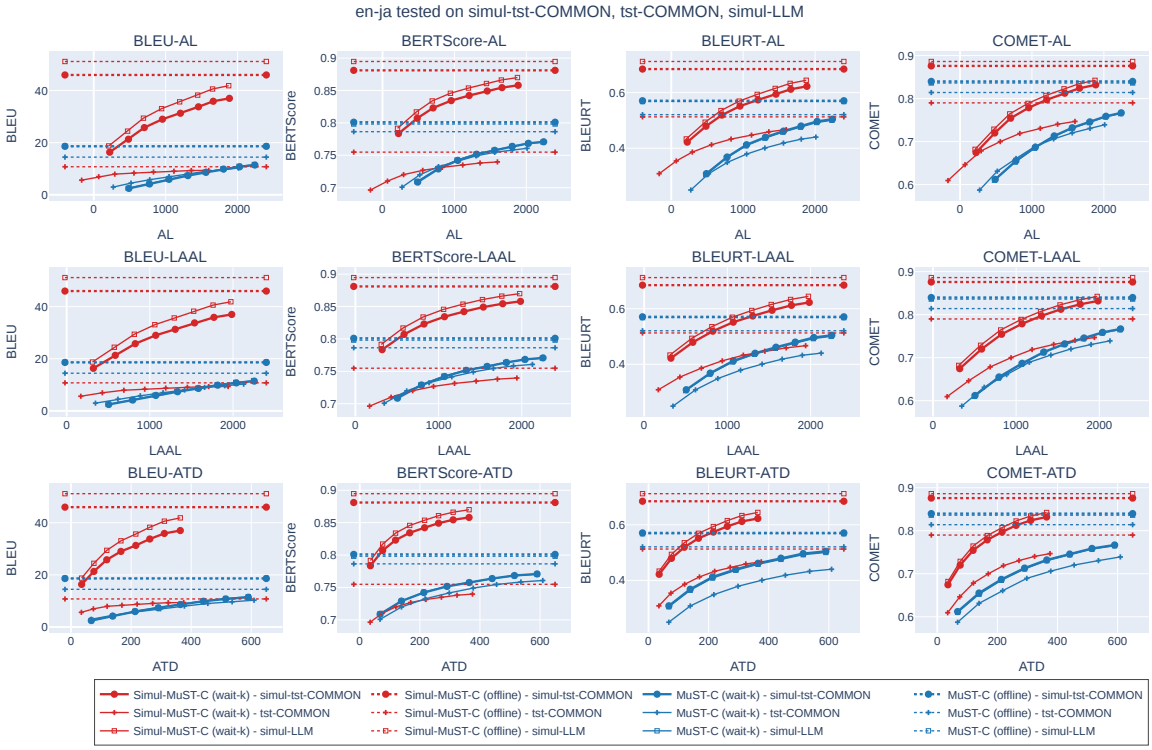

図 1　英日の同時音声翻訳結果．マーク付きの実線は wait-k, 点線はオフラインの設定を示す．

が求められるモデルの評価に適していることが示唆される．なお，英中と英独の結果については付録 7 に掲載している．

6　考察

訳出の同時性が求められるタスクで，品質と遅延のバランスを取るのは難しい課題である．実験結果から，原言語の内容を多くカバーすると品質は向上するが，出力が長くなるので出力の開始と終了を考慮する ATD のような遅延指標で不利になる．一方，出力を短くすると遅延では有利になるが，品質の低下は避けられない．実用面では，正確さが求められる場合 (例: 講義) と，スピードが求められる場合 (例: スポーツ中継) のように，異なるニーズがある．しかし，現行の評価指標はこれらを十分にサポートしていなく，今後はニーズに応じて遅延または品質いずれかの改善に焦点を当てる必要性がある．

simul-tst-COMMON を用いた評価で，英日はいずれの品質評価指標でも学習データの特徴の差が結果に現れたが，英中・英独では指標によって差が出た．具体的に，BLEU と BERTScore では，英日と同様に学習データの特徴の差が明確に現れた一方で，BLEURT や COMET ではその差が小さかった．BLEU や BERTScore は文全体でスコアを平均する前にトークンを個別に評価するが，BLEURT と COMET は一文全体を大まかに評価するため，単調性など細部の特徴を十分に捉えるのが難しい可能性がある．特に，文法的に距離が近い英独や英中にて，より細かい訳出の差を捉えるのが難しいことが予想される．そこで，単調性のように訳出スタイルが評価で重要となる場合，トークンレベルでの評価を含む指標がより適切である可能性が示唆される．

従来の研究では，同時性が求められるタスクで理想とされる単調性の程度の議論は不十分であった．本研究では，通訳者の協力を得て，SiST や SiMT に特化した評価データの作成・分析した．その結果，*simul-tst-COMMON* では，オフライン翻訳に比べて原言語文に対する目的言語の単調性が許容されることが確認できた．このテストセットは，SiST や SiMT の評価データとしての活用が期待される．

7　おわりに

本研究では，訳出の同時性が要求されるタスクに特化した評価データ *simul-tst-COMMON* を提案した．実験の結果から，*simul-tst-COMMON* は，既存のテストデータよりも適切に SiST の性能を評価できることを示した．今後の課題に，省略を含む遅延の改善に焦点を当てた SiST への取り組みが考えられる．

謝辞

本研究は JSPS 科研費 21H05054，JST 次世代研究者挑戦的研究プログラム JPMJSP2140 の助成を受けたものです.

参考文献

[1] Roderick Jones. Conference interpreting explained. In **Routledge**, 2015.

[2] Andrew Gillies. Conference interpreting: A student’s practice book. In **Routledge**, 2013.

[3] Srinivas Bangalore, et al. Real-time incremental speech-to-speech translation of dialogs. In **Proc. Conf. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 437–445, 2012.

[4] T. Fujita, et al. Simple, lexicalized choice of translation timing for simultaneous speech translation. **Proc. Interspeech**, pp. 3487–3491, 2013.

[5] Yusuke Oda, et al. Optimizing segmentation strategies for simultaneous speech translation. In **Proc. Assoc. Computational Linguistics**, pp. 551–556, 2014.

[6] He He, et al. Syntax-based rewriting for simultaneous machine translation. In **Proc. Conf. Empirical Methods in Natural Language Processing**, pp. 55–64, 2015.

[7] Yusuke Sakai, et al. Simultaneous interpretation corpus construction by large language models in distant language pair. In **Proc. Conf. Empirical Methods in Natural Language Processing**, pp. 22375–22398, 2024.

[8] Mana Makinae, et al. Simul-MuST-C: Simultaneous multilingual speech translation corpus using large language model. In **Proc. Conf. Empirical Methods in Natural Language Processing**, pp. 22185–22205, 2024.

[9] Kosuke Doi, et al. Word order in English-Japanese simultaneous interpretation: Analyses and evaluation using chunk-wise monotonic translation. In **Proc. 21st International Conference on Spoken Language Translation**, pp. 254–264, 2024.

[10] Jinming Zhao, et al. NAIST-SIC-aligned: An aligned English-Japanese simultaneous interpretation corpus. In **Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 12046–12052, 2024.

[11] Jinming Zhao, et al. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. In **Proc. Conf. Empirical Methods in Natural Language Processing**, pp. 6707–6715, 2021.

[12] Shira Wein, et al. Barriers to effective evaluation of simultaneous interpretation. In **Findings of the Assoc. Computational Linguistics: EACL 2024**, pp. 209–219, 2024.

[13] Ryo Fukuda, et al. Test data creation in simultaneous machine translation in english to japanese pair: Insights from simultaneous interpretation tactics. **IPSJ SIG Technical Report**, 2024. (In Japanese).

[14] Yuki Okamura, et al. Jyun okuri yaku” no kihan to mohan doji tsuyaku wo mohan tosita kyoikuron no shiron. In Hiroyuki Ishizuka, editor, **Word Order in English-Japanese Interpreting and Translation: The History, Theory and Practice of Progressive Translation**, pp. 217–250. Hitsuji Syobo, 2023.

[15] Yi Ren, et al. SimulSpeech: End-to-end simultaneous speech to text translation. In **Proc. Assoc. Computational Linguistics**, pp. 3787–3796, 2020.

[16] Mingbo Ma, et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In **Proc. Assoc. Computational Linguistics**, pp. 3025–3036, 2019.

[17] Danni Liu, et al. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In **Proc. Interspeech**, pp. 3620–3624, 2020.

[18] Yuka Ko, et al. Tagged end-to-end simultaneous speech translation training using simultaneous interpretation data. In **Proc. 20th International Conference on Spoken Language Translation**, pp. 363–375, 2023.

[19] Mattia A. Di Gangi, et al. MuST-C: a Multilingual Speech Translation Corpus. In **Proc. Conf. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2012–2017, 2019.

[20] Ibrahim Said Ahmad, et al. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In **Proc. 21st International Conference on Spoken Language Translation**, pp. 1–11, 2024.

[21] Alec Radford, et al. Robust speech recognition via large-scale weak supervision. In **Proc. 40th International Conference on Machine Learning**, ICML’23. JMLR.org, 2023.

[22] OpenAI, Josh Achiam, et al. Gpt-4 technical report, 2024.

[23] Zi-Yi Dou, et al. Word alignment by fine-tuning embeddings on parallel corpora. In **Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2112–2128, 2021.

[24] Kishore Papineni, et al. Bleu: a method for automatic evaluation of machine translation. In **Proc. Assoc. Computational Linguistics**, pp. 311–318, 2002.

[25] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proc. Tenth Workshop on Statistical Machine Translation**, pp. 392–395, 2015.

[26] Matthew Snover, et al. A study of translation edit rate with targeted human annotation. In **Proc. 7th Conference of the Association for Machine Translation in the Americas**, pp. 223–231, 2006.

[27] Ryo Fukuda, et al. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In **Proc. 20th International Conference on Spoken Language Translation**, pp. 330–340, 2023.

[28] Myle Ott, et al. fairseq: A fast, extensible toolkit for sequence modeling. In **Proc. Conf. the North American Chapter of the Association for Computational Linguistics**, pp. 48–53, 2019.

[29] Ashish Vaswani, et al. Attention is all you need. In **Neural Information Processing Systems**, 2017.

[30] Wei-Ning Hsu, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 29, pp. 3451–3460, 2021.

[31] Yuqing Tang, et al. Multilingual translation from denoising pre-training. In **Findings of the Assoc. Computational Linguistics: ACL-IJCNLP 2021**, pp. 3450–3466, 2021.

[32] Tianyi Zhang, et al. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[33] Thibault Sellam, et al. BLEURT: Learning robust metrics for text generation. In **Proc. Assoc. Computational Linguistics**, pp. 7881–7892, 2020.

[34] Ricardo Rei, et al. COMET: A neural framework for MT evaluation. In **Proc. Conf. Empirical Methods in Natural Language Processing**, pp. 2685–2702, 2020.

[35] Sara Papi, et al. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In **Proc. Third Workshop on Automatic Simultaneous Translation**, pp. 12–17, 2022.

[36] Yasumasa Kano, et al. Average Token Delay: A Latency Metric for Simultaneous Translation. In **Proc. Interspeech**, pp. 4469–4473, 2023.

付録

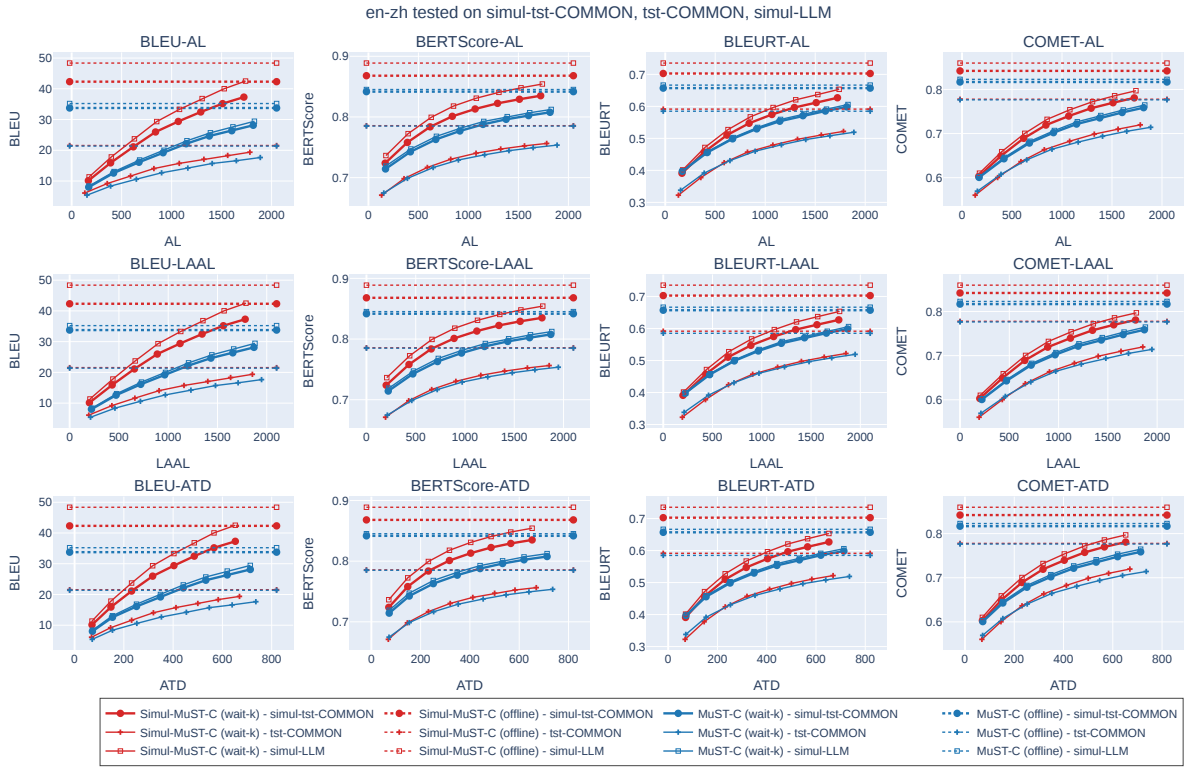

図 2　英中の同時音声翻訳結果．マーク付きの実線は wait-k, 点線はオフラインの設定を示す．

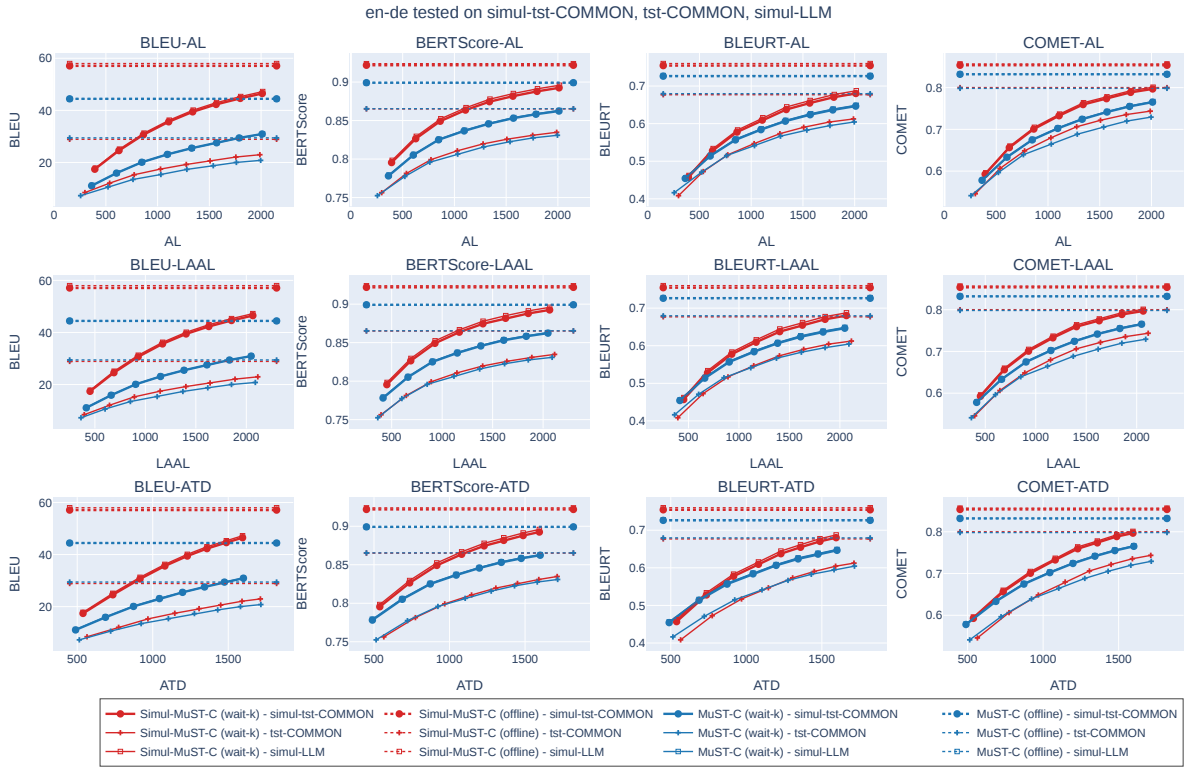

図 3　英独の同時音声翻訳結果．マーク付きの実線は wait-k, 点線はオフラインの設定を示す．