日本語講義音声のリアルタイム分割・英訳手法の評価

西村信頼[1] 武藤直輝[3] 梶田楓斗[1] 宇津呂武仁[2] レオ チーシャン[3] 西崎博光[3]
[1]筑波大学 理工学群 工学システム学類
[2]筑波大学 システム情報系 知能機能工学域 [3]山梨大学 大学院医工農学総合教育部
{s2311348,s2211268}@_u.tsukuba.ac.jp utsuro@_iit.tsukuba.ac.jp
naoki_m@_alps-lab.org {leow,hnishi}@_yamanashi.ac.jp

概要

本論文は，カスケード方式の同時音声翻訳において，句点，バッファー最大長，VAD，並びにこれら 3 手法の音声分割手法を同時に用いた際の，翻訳精度および英語字幕としての可読性に与える影響を評価した．音声認識では Whisper[1]，翻訳では DeepL を使用した．評価指標として，音声認識精度や翻訳精度の評価で広く使われる COMET[2] に加え，埋め込みベクトルの Cosine 類似度や分割適合率・再現率・F1 を用いた．評価の結果，音声認識精度が高いだけでは，字幕に適切な翻訳を生成できず，分割 F1 などで表される分割精度も重要であることを示した．

1 はじめに

近年グローバル化に伴い，日本への外国人留学生の数が増加傾向にある．そのような状況にもかかわらず，日本の大学の講義は，依然として日本語で行われているものが大半を占める．日本語は，ひらがな，カタカナ，漢字の 3 種類の文字形態を有するため，世界で最も難しい言語の 1 つと言われている．また，講義中に多数の専門用語が使用されることも，留学生の講義理解の妨げになっている．

この課題の解決策として，留学生が自分で翻訳するということが考えられる．近年は機械翻訳の研究が進み，機械翻訳の精度が向上している．しかし，翻訳には多少の時間を要するため，理解が追い付かないまま講義が進んでしまうという問題が生じる．また，教授が事前に講義の原稿を翻訳するという案も考えられる．しかし，その作業は教授側の重い負担となる．さらに，原稿を読み上げるだけでは講義は退屈なものになってしまう．それゆえ，学生の興味を引くためには，説明の合間に余談を挟むなど即興性のある講義をすることが求められるが，それを予め翻訳しておくと即興性と矛盾してしまう．

以上から，同時音声翻訳で作成した翻訳文を留学生に提供し，これらの問題を解決する．同時音声翻訳には，まず音声を認識して，その認識結果を翻訳するという手法 (カスケード方式) がある．この手法では，認識した音声を意味のある塊で区切って，その塊を翻訳して提供することで，同時性を保つ．しかし現状は，日本語で認識した音声を適度な長さで区切ることができていない．翻訳結果が長すぎることで可読性が低下し，また，講義と字幕の同時性が損なわれる．

そこで本論文では，複数の音声認識分割手法によって生成された認識結果を対象に，翻訳に必要十分な単位で音声が分割されているかを評価する．具体的には，分割位置の適切さを定量的に評価するとともに，その分割が翻訳品質にどのような影響を与えるかを分析する．

2 関連研究

同時音声翻訳の研究では，遅延を抑制するための様々な分割手法が提案されている．Ma らは，固定トークン数が入力されたら翻訳するというのを繰り返す手法 (Wait-k)[3] を提案した．また，Bangalore らが提案した無音区間で分割手法 [4] や，Fujita らが提案した右確率を用いた分割手法 [5]，清水らが提案した品詞タグを用いた分割手法 [6] 等がある。

さらに機械翻訳の評価には，n-gram の一致率を測る BLEU[7] や，単語の一致だけでなく意味の一致も考慮した COMET[2] が広く使われている．同時翻訳特有の評価では，Ma らの研究 [3] で提案された，理想遅延から何トークン遅れているのかを評価する Average Lagging(AL) や，AL をモデルの学習に利用できるように改良した Differentiable Average Lagging[8] がある．

しかし、これらの評価手法は翻訳結果全体の翻訳精度や訳出遅延に焦点を当てており，音声認識の分

割手法が翻訳精度に与える影響に関する研究は少ない．本論文は，この点に着目し，複数の音声認識分割手法によって得られる分割位置の違いと翻訳精度との関係を扱う．

3　音声認識

音声認識には，OpenAI の whisper-large-v3[1] を使用した．Whisper は Transformer ベースのエンコーダ・デコーダモデルであり，多言語音声に対して高精度な認識が可能であることから，本実験において採用した．

4　音声分割

本論文では，句点による分割，バッファー最大長による分割，VAD による分割，さらにこれらを組み合わせた計 4 種類の手法を用いた．

4.1　句点による分割

句点による分割では，音声認識に含まれる句点「。」を利用する．具体的には，Whisper の認識結果の末尾に句点「。」が出力された場合，それまでの文章を 1 分割として出力する仕組みである．句点「。」は，日本語の文末表現として使われる記号であり，これにより文のまとまりを捉えた分割になる．

4.2　バッファー最大長による分割

バッファー最大長による分割では，予め認識する固定秒数のバッファー長を定め，音声がその長さに達した場合，これまでの音声認識を 1 分割として出力する分割手法である．Whisper では最大 30 秒の音声を一度に推論することが可能であるが，実際の講義では 30 秒も発話し続けることは稀である．また，音声字幕として提示する際の可読性や文字数の制限を考慮すると，長い分割は適切ではないため，本論文では最大 15 秒のバッファーを使用する．

4.3　VAD による分割

VAD(Voice Activity Detection) による分割では，無音区間を検出し，分割を行うという手法である．本論文では，様々な言語や環境の音声データで事前学習した VAD である Silero VAD [9] を用いた．

具体的な方法は，予め無音かどうかを検出する時間長を決め，その区間において無音と判定された場合は，そこでまでの音声を 1 分割として出力する．無音区間の長さには，1 秒または 2 秒を採用した．

無音区間判定のための閾値として，1 秒の場合は 0.5，0.6，0.7，2 秒の場合は 0.7，0.8，0.9，つまり計 6 通りの閾値を採用した．これは前実験で，人目で見てよい分割を行っていると判断した閾値である．

4.4　3 手法を用いた分割

さらに，本論文では，句点による分割，バッファー最大長による分割，および VAD による無音区間に基づく分割，全て考慮した分割手法も採用した．この手法では，3 手法のいずれかの条件が満たされた時点で分割を行う．3 手法を併用した場合の分割結果が，翻訳精度にどのくらい影響を与えるか検討するため，評価対象の 1 つとした．

5　日英機械翻訳

音声認識結果を 1 分割ずつ翻訳する際には，DeepL API[1)] を使用した．DeepL は多言語に対応した，高精度ニューラル機械翻訳システムである．

本論文では，日本語講義音声を対象として留学生の授業理解支援を最終目標としている．日本語の同時翻訳では大規模言語モデル (LLM) による翻訳も考えられるが，翻訳における遅延が大きく，リアルタイム性が欠けてしまうため，低遅延かつ高精度な日英翻訳が可能である DeepL を用いた．

6　評価

6.1　音声認識

音声認識の評価手法としては，音声認識精度，分割適合率・再現率・F1 を採用した．

音声認識精度は，文字誤り率 (Charactor Error Rate) に基づいて算出した．具体的には，音声認識精度 (%) = $(1 - CER) \times 100$ とした．

分割適合率・再現率・F1 は音声認識が，翻訳に適した位置で分割できているかを評価するために採用した．評価では，まず ja-mecab を用いて音声認識正解データをトークン化し，翻訳のまとまりを考慮した分割位置を人手で作成し，正解データとした．例えば，条件や理由を表す文は，英語で If 文や Because 節として 1 文にまとめて翻訳することが自然であるため，これらを 1 分割とするように分割位置を設定した．

また，評価の際に，「次に」や「つまり」などの接続詞や，文末の「ます。」等の翻訳精度にあまり影響

1） https://www.deepl.com/ja/pro-api

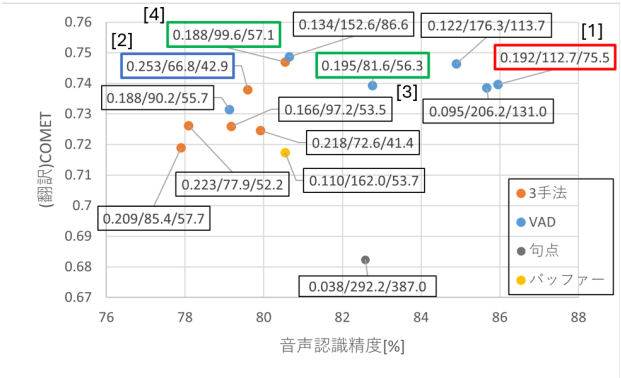

図 1　音声認識精度と COMET の関係 (ラベルは，分割 F1/翻訳文字数平均/翻訳文字数標準偏差を表す．赤枠は音声認識精度が最大，青枠は分割 F1 が最大，緑枠は文字数平均が短く，翻訳精度が高いデータを示す．)

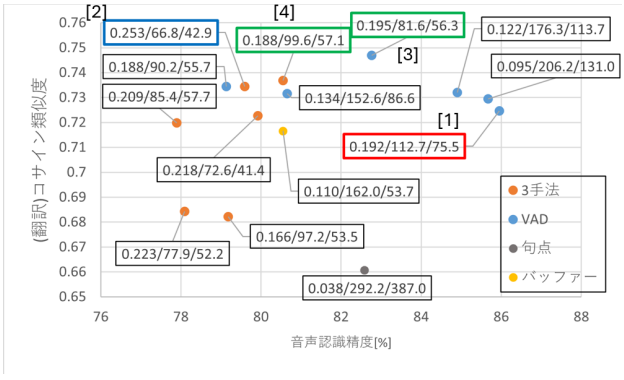

図 2　音声認識精度と埋め込みベクトルの Cosine 類似度の関係 (ラベルは，分割 F1/翻訳文字数平均/翻訳文字数標準偏差を表す．赤枠は音声認識精度が最大，青枠は分割 F1 が最大，緑枠は文字数平均が短く，翻訳精度が高いデータを示す．)

表 1　音声認識精度最大・分割 F 1 最大・翻訳後の平均文字数が短く，翻訳精度が高いデータの比較 (各評価指標で比較した時に，最も高いスコアを**太字・下線**で示す．)

ID	手法	音声認識精度 (%)	分割 F1	COMET	Cosine 類似度	文字数平均
1	VAD・無音 1 秒・閾値 0.6	**85.95**	0.192	0.7369	0.7246	112.7
2	3 手法・無音 1 秒・閾値 0.7	79.59	**0.253**	0.7379	0.7344	**66.8**
3	VAD・無音 1 秒・閾値 0.7	82.76	0.195	0.7392	**0.7470**	81.6
4	3 手法・無音 2 秒・閾値 0.8	80.55	0.188	**0.7470**	0.7369	99.6

しないと考えられる箇所は，分割位置が誤りであった場合でも正解として扱った．

本論文の，各音声の分割手法により生成された音声認識の評価結果を付録 A の表 2 に示す．

6.2　翻訳結果

翻訳精度評価には，COMET，埋め込みベクトルの Cosine 類似度，文字数平均を採用した．

COMET[2] は，原文，翻訳文，参照文を入力として，翻訳文と参照文の意味的一致率を測り，機械翻訳の精度を評価するための指標の 1 つである．本論文では，原文として音声認識結果の正解データを，翻訳文として同時音声翻訳の結果を，参照文として同時音声翻訳の正解データを使用した．このように COMET を測ることで，高精度な翻訳が可能である DeepL で翻訳をしたときに，音声認識の結果に認識誤りや不適切な分割がある場合，それらが翻訳結果に反映されてしまい，スコアが低くなると考えられる．このことから，COMET は分割手法の違いによる翻訳精度への影響を測る指標として適切である．

埋め込みベクトルの Cosine 類似度は，翻訳結果と参照文の意味的類似度を測るために採用した．埋め込みを行う際には，多言語に対応した Sentence Transformer である paraphrase-multilingual-MiniLM-L12-v2[10] を使用した．

また本論文は，日本語の講義音声の英語翻訳を字幕として，留学生に提供することを目標としているため，翻訳結果の文字数平均についても着目した．文字数の基準としては，TED が推奨する字幕文字数 (1 行 42 文字，最大 2 行)[2)] を参考にした．

機械翻訳の精度を測る指標で広く使われている BLEU[7] についても，前実験で検討した．しかし，BLEU の n-gram 一致率を測るという特性上，同時音声翻訳での言い換えや要約を適切に評価できず，本論文では適さないと考えた．

本論文の，各音声の分割手法により生成された翻訳結果の評価結果を付録 A の表 3 に示す．

これらの結果から，音声認識精度と COMET・埋め込みベクトルの Cosine 類似度，分割 F1，文字数平均の関係を図 1，2 に示す．これらの図から VAD のみの分割手法では，翻訳結果が高くなる一方で，文字数平均が長くなる傾向があることが分かる．また，3 手法を用いた分割手法では，音声認識精度が低下するものの，分割 F1 が高く，文字数平均が短くなる傾向にあることが分かる．

各指標で最も良かった分割手法の比較を表 1 に示す．この表から，音声認識精度が最大となる手法

2)　https://www.ted.com/participate/translate/subtitling-tips

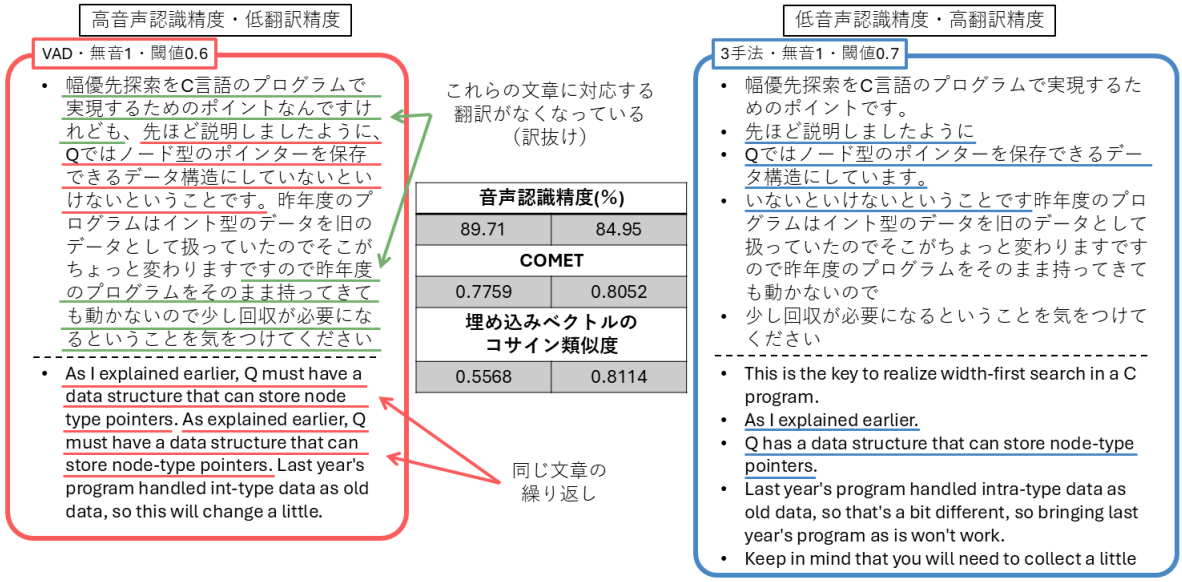

図 3　左図・赤枠: 高音声認識精度・低翻訳精度，および，右図・青枠: 低音声認識精度・高翻訳精度

(VAD・無音 1 秒・閾値 0.6) は，意味を考慮した翻訳精度評価指標である COMET や埋め込みベクトルの Cosine 類似度のスコアは，必ずしも高くなく．ま文字数平均も最も長くなっていることが分かる．

一方で，分割 F1 が最大の手法 (3 手法・無音 1 秒・閾値 0.7) と埋め込みベクトルの Cosine 類似度が最大の手法 (VAD・無音 1 秒・閾値 0.7) は，翻訳精度はともに高く，文字数平均も TED が推奨する 1 画面における字幕文字数の制限以下に収まっている．

COMET が最大の手法 (3 手法・無音 2 秒・閾値 0.8) も，文字数平均は長くなってしまっているものの，翻訳精度は高いことが分かる．

音声認識精度が高いだけでは，翻訳精度が高くならず，字幕としての可読性が低くなるという例を図 3 に示す．左図の例は表 1 で音声認識精度が最大であった分割手法，右図は分割 F1 が最大であった分割手法で生成されたものである．赤枠のものは音声認識精度が高くなっているものの，認識結果が長くなり，結果として訳抜けや，同じ文が複数回翻訳されている．一方で，青枠のものは音声認識精度が低くなっているが，適度に音声認識の分割が行われており，訳抜け等が起こっていないため，結果として COMET 及び埋め込みベクトルの Cosine 類似度のスコアが高くなった．

7　おわりに

本論文では，句点，バッファー最大長，VAD による無音区間，及びこれら 3 手法全てを用いた分割手法によって生成された分割と，それに基づく翻訳結果を評価することにより，分割手法の違いによる翻訳結果への影響について検討した。

その結果，VAD のみによる分割手法では翻訳精度が高くなるが，翻訳結果の平均文字数が長くなる傾向がある一方で，3 手法を用いた分割手法では，音声認識精度が低くなるが，分割 F1 が高くなり，字幕として表示する場合の文字数の観点での可読性が高くなる傾向があることが分かった．

本論文では，翻訳精度及び可読性に焦点を当てたが，同時翻訳では翻訳が出力するまでの遅延も大きな問題となる．そのため，今後は翻訳精度，可読性，遅延を総合的に評価することを計画している．

謝辞

本論文は，一部，科研費 25H00566 の支援を受けたものである．

参考文献

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

[2] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[3] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. Real-time incremental speech-to-speech translation of dialogs. In Eric Fosler-Lussier, Ellen Riloff, and Srinivas Bangalore, editors, **Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 437–445, Montréal, Canada, June 2012. Association for Computational Linguistics.

[5] Tomoki Fujita, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Simple, lexicalized choice of translation timing for simultaneous speech translation. In **Interspeech**, 2013.

[6] 清水宏晃, Graham Neubig, Sakriani Sakti, 戸田智基, 中村哲. 同時通訳データを利用した同時音声翻訳のための訳出タイミング決定手法. 言語処理学会第 20 回年次大会 (NLP2014), pp. 294–297, 2014.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[8] N. Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. Monotonic infinite lookback attention for simultaneous machine translation. In **Annual Meeting of the Association for Computational Linguistics**, 2019.

[9] Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

A　音声認識・音声分割・日英機械翻訳の評価結果

本節には，各分割手法における音声認識・音声分割，日英機械翻訳の評価結果を示す．表 2 は音声認識・音声分割について 6.1 節での評価指標に基づいて評価した表であり，表 3 は日英機械翻訳について 6.2 節での評価指標に基づいて評価した表である．

表 2　音声認識・音声分割の評価結果

(a) 音声分割手法: 句点・バッファー長

評価手法	句点	バッファー長
音声認識精度 (%)	82.58	80.55
分割適合率	0.556	0.667
分割再現率	0.041	0.132
分割 F1	0.038	0.110

(b) 音声分割手法: VAD (表 1 掲載分を**太字・下線**で示す)

	VAD のパラメータ					
無音認識間隔	1 秒			2 秒		
評価手法	0.5	**0.6**	**0.7**	0.7	0.8	0.9
音声認識精度 (%)	84.90	**85.95**	**82.76**	85.67	80.65	79.13
分割適合率	0.795	**0.908**	**0.695**	0.867	0.813	0.685
分割再現率	0.144	**0.243**	**0.272**	0.107	0.160	0.259
分割 F1	0.122	**0.192**	**0.195**	0.095	0.134	0.188

(c) 音声分割手法: 3 手法 (表 1 掲載分を**太字・下線**で示す)

	VAD のパラメータ					
無音認識間隔	1 秒			2 秒		
評価手法	0.5	0.6	**0.7**	0.7	**0.8**	0.9
音声認識精度 (%)	77.9	78.09	**79.59**	79.18	**80.55**	79.92
分割適合率	0.740	0.784	**0.760**	0.688	**0.753**	0.684
分割再現率	0.292	0.317	**0.379**	0.218	**0.251**	0.321
分割 F1	0.209	0.223	**0.253**	0.166	**0.188**	0.218

表 3　日英機械翻訳の評価結果

(a) 音声分割手法: 句点・バッファー長

評価手法	句点	バッファー長
COMET	0.6824	0.7173
埋め込みベクトルの Cosine 類似度	0.6608	0.7166
文字数平均	292.2	162.0

(b) 音声分割手法: VAD (表 1 掲載分を**太字・下線**で示す)

	VAD のパラメータ					
無音認識間隔	1 秒			2 秒		
評価手法	0.5	**0.6**	**0.7**	0.7	0.8	0.9
COMET	0.7464	**0.7396**	**0.7392**	0.7385	0.7487	0.7314
埋め込みベクトルの Cosine 類似度	0.7320	**0.7246**	**0.7470**	0.7295	0.7316	0.7345
文字数平均	176.3	**112.7**	**81.6**	206.2	152.6	90.2

(c) 音声分割手法: 3 手法 (表 1 掲載分を**太字・下線**で示す)

	VAD のパラメータ					
無音認識間隔	1 秒			2 秒		
評価手法	0.5	0.6	**0.7**	0.7	**0.8**	0.9
BLEU	23.7	21.9	**22.0**	26.2	**27.1**	21.3
COMET	0.7189	0.7261	**0.7379**	0.7259	**0.7470**	0.7246
埋め込みベクトルの Cosine 類似度	0.7198	0.6843	**0.7344**	0.6822	**0.7369**	0.7227
文字数平均	85.4	77.9	**66.8**	97.2	**99.6**	72.6