漸進的な音声認識・機械翻訳・テキスト音声合成に基づく音声から音声への同時翻訳

須藤 克仁　Sashi Novitasari　帖佐 克己　柳田 智也　二又 航介　Sakriani Sakti　中村 哲
奈良先端科学技術大学院大学

sudoh@is.naist.jp

1　はじめに

音声から音声への自動翻訳は異なる言語を話す人々のコミュニケーション支援における重要な課題であり，1980 年代からその基盤となる音声認識・機械翻訳・テキスト音声合成の技術と合わせ，統合システムの研究開発が続けられてきた．近年の深層学習技術によりこれらの技術は著しい進展を遂げ，発話単位の逐次翻訳については様々な商用製品が展開されるまでに至った．

一方で，同時通訳のように発話の終了を待たない漸進的な自動翻訳（同時翻訳）の試みは 2010 年代初頭に始まったが [1, 2]，機械翻訳の要素技術研究が主であり，音声入出力部分の漸進的処理を含む同時音声翻訳システムの研究はまだ十分に進んでいない．我々は，漸進的な音声言語処理に基づく同時音声翻訳システムの実現に向けた要素技術や統合システムの研究に加え，通訳者による同時通訳の技術や知見をこのシステムで活用するための同時通訳データ収集を行っている．本稿ではこれらの我々の取り組みについて紹介する．

2　自動同時翻訳の課題

2.1　同時翻訳

同時通訳 (Simultaneous Interpretation) は，通訳対象の発話を聞き取りながら，すでに聞き取った発話を別の言語へ通訳し発声する，という，非常に高度な専門技能を必要とする課題である．翻訳が主に文字で書かれた静的な入力を対象として時間をかけて精緻に訳文構成を行うものなのに対し，通訳は話者が発する逐次的な入力を対象として限られた時間の中で適宜情報を補完・要約しながら訳出を行うものである．

本研究では，入力音声に対する漸進的な処理に基づく音声から音声への同時音声翻訳 (Simultaneous Translation) に取り組む．なお，本課題では情報の補完や要約を原則行わないものとする．

2.2　同時通訳における遅延と「順送りの訳」

同時通訳に関する文献 [3] で重要な課題として挙げられているのが，通訳者の記憶や訳出の負荷に大きな影響を与えるとともに，言語間の統語構造の違いによって必然的に生じ得る訳文構成上の遅延である．

文献では以下の英文を日本語へ通訳する例が挙げられている（括弧つき数字は説明のためのもの）．

> (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.

この文の日本語訳は以下のようになる（括弧つき数字は英文との対応を表す）．

> (1) 救援担当者は (9) 生きるための (8) 食料を求めて (7) 村を荒らし回っている (6) 大量の難民たちの (5) 世話をするための (4) 十分な食料や水，宿泊施設，医薬品が (3) 無いと (2) 言っています。

ここで，(2) の say は日本語では文末に，(9) の to stay alive は日本語では主語の直後に訳出されており，その間は英語と完全に逆順となっている．このような訳を同時通訳において実現しようとすると，通訳者は (2) の動詞を保持したまま (9) までを聞き取り，その後それらを逆順にして日本語で発話することになる．その結果，主語の訳出以降英文を聞き終わるまで通訳発話を再開できず残りのすべての情報を一時的に記憶しなければならない上に，聞き取りから発話までの遅延 (ear-voice span) が非常に大きくなり，次の発話の聞き取りと通訳に重大な支障を来す可能性が高い．そこで，同時通訳では記憶負荷や遅延を低減するために以下のような「順送りの訳」がしばしば用いられる．

(1) 救援担当者たちの (2) 話では (4) 食料，水，宿泊施設，医薬品が (3) 足りず (6) 大量の難民たちの (5) 世話ができないとのことです．(7) 難民たちは今村々を荒らし回って，(9) 生きるための (8) 食料を求めているのです．

この訳では，英文の要素を前から小分けにして訳出するとともに，(7) 以降の関係詞節の手前で一旦文を区切り，関係詞節の内容を後から補足することで日本語としての自然さを損なわないようにしている．この問題は英語と日本語で語順が大きく異なることに起因しており，順送りの訳のような工夫なしでは英語と日本語の間の同時通訳は困難であると言える．

2.3 自動同時翻訳における遅延

音声から音声への自動同時翻訳においても音声認識・機械翻訳・テキスト音声合成の各段階での遅延が問題となる．通常これらの技術は文を入力単位として実装されるため，一文の発話が終了してから順に各処理と結果の受け渡しが行われるために発話時間分の遅延が必ず発生する．話し言葉においては文の境界がしばしば不明瞭であるために処理単位が長くなり得る上，実際のシステムでは各処理における処理時間分さらに遅延が増大する．機械翻訳においては前節で述べたように語順の入れ替えのための遅延の影響が大きく，音声認識やテキスト音声合成では順序の入れ替えは不要であるものの入力文確定後に処理を行う方式ではこうした遅延は避けられない．そこで本研究では，これらの処理を漸進的に行う技術を統合して，同時翻訳における遅延の低減を目指す．

3 漸進的音声言語処理

本節では音声認識・機械翻訳・テキスト音声合成における漸進的処理のための手法について簡単に述べる．

3.1 漸進的音声認識

音声認識においても注視機構付き系列変換 (attentional sequence-to-sequence) モデルが広く用いられている．しかし，通常は文全体の状態系列を注視対象としており漸進的な処理に対応できないため，後方の文脈を参照しないような特殊なモデルや学習方法が試みられてきた [4]．

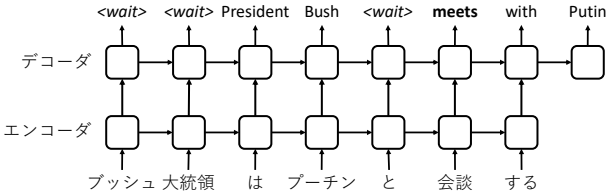

図 1: 適応的な入力待機を行う漸進的機械翻訳．

我々は，文全体を入力して注視するモデルを教師 (teacher) とし，漸進的処理のために短いセグメント単位で注視を行うモデルを生徒 (student) として，生徒が教師の注視を再現できるように音声認識の学習を行う手法を提案した [5]．遅延を最小限に留めるために各セグメントの情報のみで音声認識を行うと十分な精度が得られないため，400ms 程度の遅延を許容して対象セグメントの後方の音声特徴量も利用することで文単位の入力を利用した場合からの精度低下を抑えられることを実験的に確認した．

3.2 漸進的機械翻訳

機械翻訳では先に述べたように語順の違いにより低遅延での訳出が難しい場合がある．英日翻訳・日英翻訳における順送りの訳の学習のためのデータ量は通常の対訳データ量と比較して著しく不足しており，現在は多くの研究において（順送りではない）通常の翻訳を行った対訳コーパスから翻訳の学習を行っている．

低遅延での同時翻訳を実現する方法として提案されたのが *wait-k* と呼ばれる，入力トークン列に対して k トークンの入力を待ってから翻訳出力を開始する方式である [6]．ある時点での訳語選択に必要な情報がそれ以前の入力で得られていない場合は，それ以前の入力から強制的に訳語選択を行うこととなり，ある種の予測として機能する．*wait-k* は非常に単純な方式で実装も容易だが，英語と日本語の間のような語順の差が大きい場合には不十分である．

我々は，デコーダの出力記号にトークンを出力せず次の入力を待つことを表す特殊記号を追加し，訳語選択に必要な入力が得られていない場合に適応的に入力を待つ方式を提案した [7]．提案手法の動作を模式的に表したものを図 1 に示す．英語から日本語への翻訳実験においては，*wait-k* では十分な入力が得られず過度な予測を求められるのに対して，提案手法は適応的な入力待機を行い漸進的な翻訳による精度低下を小さく抑えられることを確認した．

3.3 漸進的テキスト音声合成

テキスト音声合成における漸進的な処理は，音声認識や機械翻訳に比べてあまり多くの検討が行われていない状況にある．系列変換モデルによる end-to-end 処理はテキスト音声合成でも活用され，合成音の自然性が大きく向上するに至っているが，自然な合成音の予測には周辺の単語から得られる特徴量が不可欠である．HMM に基づくテキスト音声合成では漸進的な処理のために特徴量の予測を組み込んだ手法が既に提案されているが [8]，系列変換モデルを用いた漸進的な処理についてはこれまで試みられていなかった．

我々は，単語（英語の場合）やアクセント句（日本語の場合）を単位として入力テキストをセグメントに分割し，セグメントごとに音響パラメータ（スペクトログラム）の予測やセグメント終端の予測を行う漸進的な end-to-end テキスト合成手法を提案した [9]．提案手法を利用した主観評価実験により，1 単語／アクセント句のみの情報に基づく漸進的音声合成は自然性が低く，多少の遅延を許容して 2-3 単語／アクセント句の情報を利用するほうが自然性が高いことを実験的に確認した．

4 音声から音声への同時翻訳システム

前節で述べた漸進的音声認識・機械翻訳・テキスト音声合成技術を統合して，音声から音声への同時翻訳を行う試作システムを作成した．本試作システムは英語の講演音声を日本語の音声に翻訳するものであり，TED Talks の英語講演の翻訳を主たる対象としている．

本試作システムは各処理モジュールを単純にカスケード接続したもので，以下のような手順で処理を行う．

- マイクもしくは音声／動画ファイル入力[1]の英語音声に対する漸進的音声認識を行い，その結果を機械翻訳モジュールに渡す．
- 音声認識モジュールから得られた英語の音声認識結果を日本語に翻訳し，その結果をテキスト音声合成モジュールに渡す．
- 機械翻訳モジュールから得られた日本語への翻訳結果を音声合成し，スピーカーもしくは音声ファイルに出力する．

[1]ファイル入力時はファイル読み込みが音声の実時間より高速である点には注意を要する．

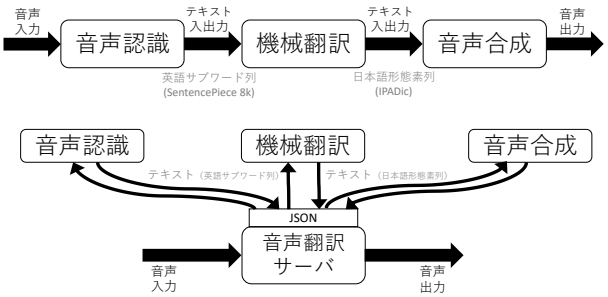

図 2: 試作システムの構成例．

なお，モジュール間での処理単位の変換を最小限に留めるため，音声認識結果が後段の機械翻訳の入力側と同じ (SentencePiece[10] による) サブワードモデルに基づいたサブワード列として得られるように音声認識モデルを学習し，機械翻訳結果が後段のテキスト音声合成で用いる IPADic に基づく日本語形態素の列の形で得られるように機械翻訳モデルを学習した．学習データはそれぞれの論文に記載のものに加え，TED Talks 英語講演と日本語字幕のデータを利用した（テキスト音声合成モデルを除く）．

本試作システムにおける各モジュール間の接続は (1) テキストによる標準入出力（パイプ接続） (2) 統括サーバとの相互通信 のいずれかで行う（図 2）．単一の入力音声に対する処理であれば，各モジュールを別プロセスで駆動した (1) の構成で動作可能である．

5 同時通訳データの収集

前節で述べた同時音声翻訳技術・システムの研究に加え，本研究のための講演同時通訳データの収集を継続的に行っている．概ね 5 年以上の通訳経験を有する通訳者を中心に[2]，対象の音声（映像を含む対象については映像も参照可能）を通訳した音声を録音し，順次書き起こしを付与している．本稿執筆時点までに，TED Talks を中心に英語から日本語で約 150 時間分，日本語から英語で約 110 時間分が収録済みである．今後も講演以外のデータも含め収集していく予定である．

また，順送りの訳でない対訳データの目的言語側の文を順送りの訳に近い文に変換する方法 [11] についても合わせて検討を行っており，上記同時通訳データの整備と合わせて同時翻訳システムの性能向上に繋げることを目指している．

[2]一部のデータについては通訳経験年数の異なる複数の通訳者の通訳を収録している．

6 課題と今後の展望

本試作システムは音声から音声への同時翻訳を志向した漸進的音声認識・機械翻訳・テキスト音声合成技術の連携によって実現したものである．同時翻訳システム全体としての性能向上のためには当然各モジュール単位での精度および処理効率の向上も重要であるが，各処理の効果的な統合方法がシステム全体として見たときの重要な課題である．現在のカスケード接続では後続の処理への誤り伝播が避けられないため，モジュール間の接続において 1-best の処理結果だけでなく n-best や単語ラティスのように曖昧性を含んだ結果を渡し，それを考慮した処理が行われることが好ましい．また，近年音声から音声への end-to-end 翻訳 [12] が注目を集めつつあり，同時翻訳においてそうしたアプローチの有用性についての検討が必要であろう．

音声から音声への同時翻訳の今後の展望として，実際の応用において自動同時翻訳がどの程度有用であるかを検証・評価することが考えられる．同時翻訳に関する研究 [2, 6] では BLEU 等の翻訳精度と遅延の大きさのトレードオフとしてその性能を議論してきたが，実際には情報の受け手にとって有用であったか，という観点での議論が必要であろう．さらに，同時通訳のように時間制約の中で情報を要約したり，外部知識や講演資料等に基づいて発話内容を予測したりする等，より通訳に近い処理の実現も将来の目標と考えることができよう．

7 おわりに

本稿では，漸進的な音声認識・機械翻訳・テキスト音声合成技術に基づく音声から音声への同時翻訳のアプローチと，その試作システムについて述べた．今後も引き続き個々の技術および統合手法のさらなる検討と同時通訳データの蓄積を行う予定である．

謝辞 本研究は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. Real-time incremental speech-to-speech translation of dialogs. In *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.

[2] Tomiki Fujita, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation. In *Proceedings of Interspeech*, pp. 3487–3491, 2013.

[3] 水野的. 同時通訳の理論—認知的制約と訳出方略. 朝日出版社, 2015.

[4] Kyuyeon Hwang and Wonyong Sung. Character-level incremental speech recognition with recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5335–5339, March 2016.

[5] Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition. In *Proceedings of Interspeech 2019*, pp. 3835–3839, 2019.

[6] 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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] 帖佐克己, 須藤克仁, 中村哲. 英日同時翻訳のための Connectionist Temporal Classification を用いたニューラル機械翻訳. 情報処理学会研究報告 2019-NL-241, 2019.

[8] Timo Baumann. Decision tree usage for incremental parametric speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3819–3823, May 2014.

[9] Tomoya Yanagita, Sakriani Sakti, and Satoshi Nakamura. Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework. In *Proceedings of the 10th ISCA Speech Synthesis Workshop*, pp. 183–188, 2019.

[10] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[11] 二又航介, 須藤克仁, 中村哲. 英日同時通訳システムのための疑似同時通訳コーパス自動生成手法の提案. 言語処理学会第 26 回年次大会発表論文集, 2020.

[12] Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. Interspeech 2017*, pp. 2630–2634, 2017.