Vol.2019-NL-242 No.4
2019/10/25

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

Sashi Novitasari[1]　帖佐 克己[1]　柳田 智也[1]　二又 航介[1]　須藤 克仁[1,a)]　Sakriani Sakti[1]　中村 哲[1]

概要：講演等の同時通訳に代表されるような漸進的な翻訳処理は，遅延の少ない音声言語コミュニケーションの実現のために有益である．音声から音声への自動翻訳が盛んになってきているが，その多くは 1 文単位の逐次的な処理に基づくものであり，音声認識，機械翻訳，音声合成の各段階で 1 文単位の入力が揃うまでの遅延が発生し，全体として大きな遅延を生むことになる．我々は，3 つの各モジュールがそれぞれ漸進的な処理を行い，少ない遅延で後段のモジュールに結果を受け渡すことで漸進的な音声から音声への自動翻訳を実現するシステムの開発を行っている．本稿では，漸進的処理を行う音声言語処理の要素技術とそれらを統合した試作システム，および今後の展望について述べる．

1. はじめに

音声言語は人間同士のコミュニケーションの中心的な役割を果たすものであり，計算機を使った音声言語処理はコミュニケーション支援の有望な技術として長きに渡り研究開発が続けられてきた．その主要な技術である音声認識，機械翻訳，テキスト音声合成はそれぞれ広範に用いられるようになり，特に昨今の深層学習技術の発達によってさらなる性能向上と適用領域の拡大が進んでいる．

これらの技術を統合して用いる音声から音声への翻訳は音声言語コミュニケーション支援技術として最も挑戦的な課題の一つである．1980 年代から始まった ATR の研究では，会議室の予約受付や旅行における会話等，比較的文が短く，語彙も限られたタスク設定において，発話単位での逐次通訳の実現可能性を示した．統計的手法による音声言語処理技術の進展，音声言語リソースの蓄積，計算機の高速化，が進むにつれ文の長さや語彙サイズの制限が徐々に緩和され，スマートフォンを使った音声逐次翻訳が実現されるに至った．一方で，同時通訳のように発話の終了を待たずに漸進的な翻訳を行う同時音声翻訳の研究が本格化したのはこの 10 年ほどである [1], [2]．そこから要素技術である音声認識・機械翻訳・テキスト音声合成が深層学習に基づく高精度なものに進化し，そうした新しい技術によって同時音声翻訳の実現に向けた取り組みが進みつつある．

本稿では，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技術を紹介するとともに，それらを統合した同時音声翻訳の試作システムについて述べる．

[1] 奈良先端科学技術大学院大学
NAIST (Nara Institute of Science and Technology), Ikoma, Nara 630–0192, Japan
[a)] sudoh@is.naist.jp

2. 自動同時翻訳の課題

2.1 通訳と翻訳

同時通訳 (Simultaneous Interpretation) は，通訳対象の発話を聞き取りながら別の言語への通訳を行い発声するという，非常に高度な専門技能を必要とするタスクである．通訳においては，「聞く」と「話す」とを次々に切り替え，それと同時に聞き取った内容を解釈して，時に補い，時に省きながら訳文を構成しなければならない．このように，連続的な発話という時間制約の中で別の言語で内容の伝達を行うことは，人間の認知・言語運用の観点で見ても非常に挑戦的な課題であると言える．書き言葉の文書に対する翻訳が静的な入力を対象としたタスクであり，前後の文脈を考慮し，時に外部資料を参照しながら精緻に訳文構成を行っていくものであるのに対し，話し言葉の発話に対する通訳は動的な入力を対象としたタスクであって，事前の資料と直前までの文脈だけを利用して，限られた時間で訳出を行わねばならないものである．また，記者会見や外交の場面等でよく見られる，発話単位で通訳を行う逐次通訳に対して，同時通訳においては通訳対象の発話は通訳の進捗と関わりなく進んでしまうため，通訳の遅延は訳出の遅れという問題のみならず通訳者による聞き取りにも深刻な影響を与えるものであって，遅延が生じないようにするための様々な工夫が行われている．

本研究では，同時通訳にみられるような情報の補完や要約については現時点では扱わないこととし，入力音声

に対する漸進的な処理に基づく同時翻訳 (Simultaneous Translation) に取り組む．

2.2 同時通訳における遅延と「順送りの訳」

同時通訳における種々の課題については文献 [3] に詳しいが，その中でも重要な課題の一つとして挙げられているのが，言語間の統語構造の違いによって必然的に生じうる訳文構成上の遅延である．文献に挙げられている例を以下に示す．

まず，次の英文を日本語に通訳することを考える（括弧つき数字は説明のために付されたものである）．

(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.

この文を日本語に訳すとすると，通常は以下のように訳文を構成するであろう（括弧つき数字は英文と対応する日本語の節や句を示す）．

(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。

一見して分かる通り，(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，その間は英語と逆順となっている．このような訳文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，そこから聞き取った内容を逆順に日本語にして発話する必要がある．こうした処理は通訳者の短期記憶に強い負荷を与えるものであるとともに，英文を聞き終わるまで通訳発話を開始できないため非常に大きな遅延を生じてしまう．そこで，以下のように訳出を行うことで記憶負荷と遅延の減少を図る「順送りの訳」がよく利用される．

(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し，(7) 以降の関係詞節はその手前で一旦文を区切って，関係詞節の内容は「難民」を補足する文として付け加え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．これは英語が主辞前置型 (head-initial) 言語，日本語が主辞後置型 (head-final) 言語であって語順が大きく異なる[*1]ことに起因する．そのため，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

ここまで述べた通り，英語と日本語の間の音声翻訳は本質的に遅延が生じやすい．さらに，計算機による音声から音声への同時翻訳を音声認識・機械翻訳・テキスト音声合成を連結して実現しようとした場合には，それぞれのプロセスにおける遅延が生じることになる．特に，それぞれのプロセスが文を入力の単位として仮定して設計されている場合には，一文の音声入力が終了してから音声認識の処理時間分の遅延が生じた上で結果が機械翻訳に渡され，さらに機械翻訳とテキスト音声合成の処理時間分の遅延が追加されて音声の出力が開始される．これは講演等文が長くなりがちな状況においては，膨大な遅延を生じさせる要因となる．

この問題を解決するための方法として，文の終了を待たずに漸進的に処理を行う技術が知られている．本研究では次節に示す我々が研究を進めている漸進的音声言語処理の技術を統合し，本節で示した処理機構上の遅延と，英語と日本語との統語構造の差によって生じる遅延の削減を目指す．

3. 漸進的音声言語処理

本節では音声から音声への翻訳システムを構成する，音声認識・機械翻訳・テキスト音声合成の各々において漸進的処理を行うための技術について簡単に述べる．技術の詳細や個々の性能評価についてはそれぞれ文献 [6], [7], [8] を参照されたい．

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が，通常注視の対象が文単位の状態系列であることから，漸進的な処理に対応できない．漸進的な処理を実現するために，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提案されている [9]．

我々の研究 [6] では，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．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 精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が実験的に確認されている．

[*1] これは統計的機械翻訳における単語の並べ替えにおいても大きな問題であり，英日翻訳において主語の後ろをすべて逆順にしたり [4]，統語解析を行い主辞後置化を行ったり [5] する方法が考案されており，機械翻訳においてもよく知られている．

3.2 漸進的機械翻訳

機械翻訳ではすでに述べたように語順の違いによって低遅延での訳出が難しい場合がある．順送りの訳の実現には依然としてデータ量が不足していることもあるため，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のが実情である．そうした状況で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* [10] 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．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．

我々の研究 [7] では後段の入力を適応的に待つ手段として，デコーダの出力記号の一つ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．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，提案手法は適応的に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が確認されている．

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従来の取り組みが少ない研究課題と言える．テキスト音声合成でも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り，漸進的な処理のために特徴量の予測を HMM に基づくテキスト音声合成に組み込んだ手法が提案されている?．ニューラルネットワークに基づく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てきているが，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．

我々の研究 [8] では，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．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音声合成よりも，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した音声合成のほうが自然性が高いことが確認されている．

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利用して，音声から音声への同時翻訳を実現する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る．

図 1 試作システムによる字幕重畳表示の例．字幕は上から手動書き起こし，音声認識結果，翻訳結果の順．

本試作システムは各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[*2]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．

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) 処理統括サーバとの相互通信 のいずれかで行う設計とした．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) の構成で動作させることが可能である．

英語の講演映像に対して本試作システムで翻訳を行い，字幕重畳表示を行った例を図 1 に示す[*3]．実際には漸進的な音声認識・機械翻訳の進行に合わせて字幕を更新し，テキスト音声合成の進行に合わせて合成音声を再生することができる．

[*2] 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．
[*3] なお，翻訳における処理遅延のため，翻訳字幕は直上の音声認識結果字幕とは対応しない．

5. 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が必須であるが，それと同時に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?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], [10] 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．

6. 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．今後は技術の検討と実際の同時通訳データの蓄積をさらに進める予定である．

謝辞 本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Bangalore, S., Rangarajan Sridhar, V. K., Kolan, P., Golipour, L. and Jimenez, A.: Real-time Incremental Speech-to-Speech Translation of Dialogs, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, Association for Computational Linguistics, pp. 437–445 (online), available from ⟨https://www.aclweb.org/anthology/N12-1048⟩ (2012).

[2] Fujita, T., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation, *Proceedings of Interspeech*, pp. 3487–3491 (online), available from ⟨https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_3487.pdf⟩ (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] Katz-Brown, J. and Collins, M.: Syntactic Reordering in Preprocessing for Japanese → English Translation: MIT System Description for NTCIR-7 Patent Translation Task, *Proceedings of the NTCIR-7 Workshop Meeting*, pp. 409–414 (2008).

[5] Isozaki, H., Sudoh, K., Tsukada, H. and Duh, K.: Head Finalization: A Simple Reordering Rule for SOV Languages, *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, Association for Computational Linguistics, pp. 244–251 (online), available from ⟨https://www.aclweb.org/anthology/W10-1736⟩ (2010).

[6] Novitasari, S., Tjandra, A., Sakti, S. and Nakamura, S.: Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition, *Proceedings of Interspeech 2019*, pp. 3835–3839 (online), DOI: 10.21437/Interspeech.2019-2985 (2019).

[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] Yanagita, T., Sakti, S. and Nakamura, S.: Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework, *Proceedings of the 10th ISCA Speech Synthesis Workshop*, pp. 183–188 (online), DOI: 10.21437/SSW.2019-33 (2019).

[9] Hwang, K. and Sung, W.: Character-level incremental speech recognition with recurrent neural networks, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5335–5339 (online), DOI: 10.1109/ICASSP.2016.7472696 (2016).

[10] Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H. and Wang, H.: STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 3025–3036 (online), DOI: 10.18653/v1/P19-1289 (2019).