Vol.2021-SLP-135 No.5
2021/2/18

次世代音声翻訳の研究（科研費基盤研究(S)）

中村　哲 [1]
s-nakamura@is.naist.jp

概要：2017 年から進めている次世代音声翻訳の研究について紹介する．この研究では，講演，講義の同時音声翻訳やパラ言語の音声翻訳の基礎研究，および同時通訳コーパスとプロトタイプシステムの構築を目指している．

キーワード：音声同時翻訳，インクリメンタル音声認識・合成・機械翻訳，パラ言語音声翻訳，同時通訳コーパス

1. はじめに

旅行会話に対する音声翻訳技術については実用化が進んでいる．しかし，講演，講義，会議のような場面での自動音声同時通訳は技術的に格段に困難で研究が進んでいない．著者は 2012 年からの基礎的なコーパス構築，要素技術研究を経て，2017 年から科研費基盤 S の支援を得て，本格的に大規模コーパス，要素技術研究，プロトタイプの構築を開始した．本稿では研究の狙い，体制，進捗について紹介する．

2. 研究の狙い

先に述べたように，短い旅行会話を対象に一発話終了毎に翻訳する音声翻訳は実用化が進んでいるが，人間の通訳者が行うような同時通訳は格段に困難である．特に文構造が異なる日本語から英語の通訳では，文末に来る動詞や否定を待つかを予測しなければ訳出ができない．本研究では，講演，講義を対象に，発話者の音声を常時音声認識し，言語間での文構造の違いを考慮して五月雨式に通訳する自動音声同時通訳と音声翻訳の高度化の研究を中心に，発話者の感情，強調，話者性等を抽出，保持，生成するパラ言語音声翻訳，講演，映像などのビデオコンテンツの字幕翻訳，音声画像翻訳，脳活動を含むセンシングによるリアルタイムコミュニケーション測定，の研究を行い，同時通訳コーパス構築とプロトタイプシステムを構築する．

3. 研究の体制

研究の体制を右図に示す．音声翻訳だけでなく，音響信号処理，音声認識，音声合成，声質変換，機械翻訳，対話処理，認知処理の専門家による研究チームとなっている．

4. 研究の進捗

紙面の都合もあり，本稿では主として奈良先端大で担当している同時通訳およびパラ言語翻訳に関する進捗について紹介する．

4.1 同時音声翻訳[1]

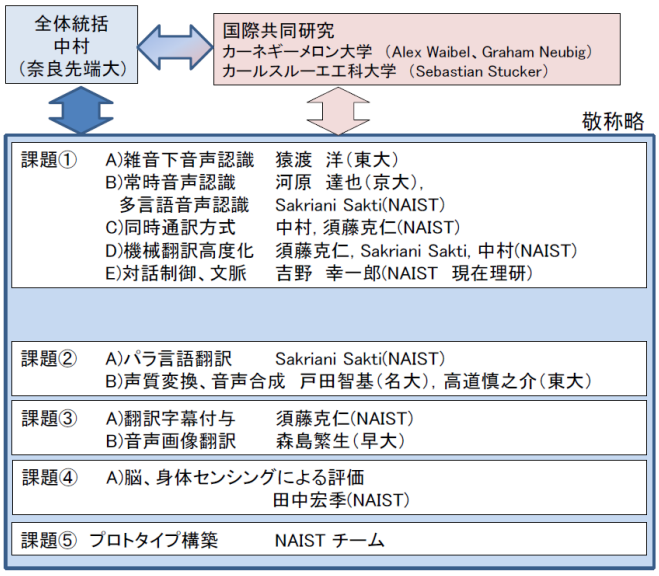
音声翻訳は，通常，音声認識，機械翻訳，テキスト音声合成から構成されるが，昨今の深層学習技術の発達によって著しい性能向上が進んでいる．しかし，同時通訳のように発話の終了を待たずに漸進的な翻訳を行う同時音声翻訳の研究が本格化したのはこの 10 年ほどである[1, 2].

人間による同時通訳は，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非常に高度な専門技能を必要とするタスクである．書き言葉の文書に対する翻訳が静的な入力を対象とし前後の文脈を考慮し，時に外部資料を参照しながら時間をかけて訳文構成を行うのに対し，同時通訳は話し言葉の動的な発話入力を対象としたタスクであり，事前の資料と直前までの文脈だけを利用し，情報を要約したりしながら，限られた時間で訳出を行う．このような情報の補完や要約を含まないシステムを本研究では同時音声翻訳システムと呼び，必要なコーパス，要素技術，プロトタイプの研究開発を行う．

同時通訳，特に日本語と英語のように語順の違う言語の同時通訳の問題は非常に困難とされている．本節では同時音声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．（詳細は文献[3, 4, 5]）

4.1.1 漸進的音声認識

音声認識では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，双方向 LSTM



1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology.

等が使われていて，注視の対象が文全体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．我々の研究[3]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．400ms 精度の遅延を許容して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．

4.1.2 漸進的機械翻訳

機械翻訳では語順の違いによって低遅延での訳出が難しい．現在は通常の対訳コーパスから翻訳モデルの学習を行って同時音声翻訳の研究を行っている．低遅延での同時翻訳を実現するために提案された方法の一つが wait-k [6]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．ただ，wait-k は，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[4]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，提案手法は適応的に入力待機を行い wait-k に比べて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．

4.1.3 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については，HMM ベースのテキスト音声合成に組み込んだ手法が提案されている[7]．我々は，ニューラルネットワークに基づく end-to-end テキスト音声合成が漸進的に動作するシステムを開発した[5]．この方法では，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．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，自然性が高いことを確認した．

4.1.4 同時通訳データの収集

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．本稿執筆時点までに，TED Talks を中心に英語から日本語で約 130 時間，日本語から英語で約 130 時間の熟練の通訳者による同時通訳を収集し，原言語のアノテーション，通訳者音声の録音，アノテーションをしており，今後も TED 等の講演以外のデータも含めコーパス構築を継続していく予定である．

4.2 パラ言語音声翻訳

音声から音声への音声翻訳では，入力発話における強調や感情などのパラ言語情報を出力発話に付与することがコミュニケーションを成立させるために重要である．我々の研究では，入力音声から平常発話と強調発話から学習された回帰 HMM を用意しておき，入力発話の強調度合いを抽出する. 音声認識の結果と強調度合いの系列を, それぞれ，エンコーダ・デコーダによるテキスト翻訳と条件付き確率場に基づく強調度合い変換により変換し，目的言語で音声を合成する[8]．さらに，LSTM に基づくエンコーダ・デコーダモデルで，テキスト翻訳と強調度合い翻訳の両方を同時に変換する研究を行っている．この方法を適用した音声の主観評価実験を行ったところ，83%の割合で強調を聴取できることが明らかとなった[9].

単語やフレーズの強調は音声による強調だけでなく文としての強調も可能である．我々はこの検討のために，強調の度合いを音声で変化させたもの，文で変化させたものを用意し音声翻訳後の強調がどのように表現できるかについて検討を行った．5 段階の強調度合いを含む文を音声合成し等価性を主観評価したところ，概ね等価関係は維持されるが文表現に対し音声表現の強調の方が強調度合いを明確に表現できることが明らかとなった[10].

謝辞

本基盤 S プロジェクトの共同研究者である，東京大学猿渡先生，高道先生，早稲田大学森島先生，名古屋大学戸田先生，京都大学河原先生，本学の須藤先生，Sakti 先生，田中先生, 吉野先生(現理研), および学生諸君に感謝します．本研究は JSPS 科研費 JP17H06101 の助成を受けました．

参考文献

[1] Srinivas Bangalore, et al., “Real-time incremental speech-to-speech translation of dialogs”. NAACL 2012, pp. 437-445,
[2] Tomoki Fujita, et al., “Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation”, Interspeech, pp. 3487-3491, 2013.
[3] Sashi Novitasari, et al., “Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition”, Interspeech 2019, pp. 3835{3839, 2019.
[4] 帖佐克己, 他, “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.
[5] Tomoya Yanagita, et al., “Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework”, ISCA Speech Synthesis Workshop, pp.183{188, 2019.
[6] Mingbo Ma, et al., “STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework”. ACL 2019 pp. 3025-3036,
[7] Timo Baumann, “Decision tree usage for incremental parametric speech synthesis”, IEEE ICASSP 2014 pp.3819-3823
[8] Quoc Truong Do, et al., "Preserving Word-level Emphasis in Speech-to-speech Translation", IEEE/ACM TASLP25, 3, 544-556, Dec. 2016
[9] Quoc Truong Do, et al, "Toward Expressive Speech Translation: A Unified Sequence-to-Sequence LSTMs Approach for Translating Words and Emphasis", Proc. INTERSPEECH, Aug. 2017
[10] Quoc Truong Do, et al., “Toward Multi-features Emphasis Speech Translation: Assessment of Human Emphasis Production and Perception with Speech and Text Clues”, IEEE Spoken Language Technology Workshop (SLT), Dec. 2018