Vol.2019-MUS-123 No.2
Vol.2019-SLP-127 No.2
2019/6/22

自動音声翻訳から自動音声通訳へ

中村　哲[†1]

概要：音声翻訳は入力音声をその場で認識，翻訳，合成する技術であり，近年の Sequence-to-sequence 型のニューラルネットワークの登場により大きな進歩を遂げている．本稿では，これらの技術について紹介する．

Toward Speech Interpretation from Speech Translation

Satoshi Nakamura[†1]

1. はじめに

近年の深層学習，系列モデリング技術の進歩により音声認識，音声合成は大きな進歩を遂げた．一方，自然言語処理でも潜在空間への埋め込み，ニューラルネット等による系列モデリング技術により，連続空間で処理が再定義され，多くの問題で著しい進歩が見られた．このことは，音声処理と自然言語処理を統合した音声言語処理をより一貫した形でできる時代が到来したことを示している．本稿では，音声翻訳研究について，現状と今後の方向について述べる．音声翻訳については，これまでも著書や解説記事で紹介してきたので，詳細はそちらを参照されたい [1] [2] [3]．音声認識では HMM 黎明期，第 2 期ニューラルネットワークの登場，そして HMM＋N-gram＋WFST が定着し，そして，2000 年後半から急激に深層学習，強化学習を中心とする第 3 期ニューラルネットワークの時代に入った．その結果，Switchboard や Call home 自由発話タスクで人間の書き起こし性能に匹敵する音声認識性能が達成されるまでに進歩した．一方，機械翻訳ではルールベース翻訳，フレーズベース統計翻訳，そして，2010 年代に LSTM(Long Short Term Memory)をはじめとするリカレントニューラルネットワークによる機械翻訳が登場し，飛躍的な性能改善をもたらした．

2. 音声処理，自然言語処理の動向

本章では，音声言語処理に入る前に，音声処理，機械翻訳のそれぞれの最近の進歩について紹介する．

2.1 Sequence-to-sequence 型音声認識・合成

深層学習に基づく音声認識の最近の動向としては，HMM の音響モデル確率を DNN の事後確率に置き換える DNN-HMM が基本である．しかし，近年では Sequence-to-sequence の音声認識手法が注目されている [4]．この方法では，入力のパラメータ系列を LSTM でモデリングし直接文字列を出力するように学習する．音声合成も，Oden らにより Wavenet が提案され [5]，音響モデルが波形ベースで学習できるようになった．さらに Wang らは [6] で，Sequence-to-sequence の考え方で音声合成システム Tacotron を開発した．入力は文字列で one-hot ベクトルあり，それが Sequence-to -sequence の注視機構付きニューラルネットワークに入力され，最終的にスペクトルを生成した後，Griffin-Lim アルゴリズムで波形を生成する．主観評価を行ったところ素片接続（MOS:4.09）よりは低いものの，良好な評価(MOS:3.82)を得ている．また，Tacotron の波形の生成に Griffin-Lim ではなく，Wavenet Vocoder を用いた Tacotron2 も開発されている [7]．

2.2 機械翻訳の動向

機械翻訳についても，2014 年頃まではフレーズベースの統計翻訳，また，構文構造を確率的に推定しながら翻訳を行う Tree-to-string, Forest-to-string の研究が主流であった．一方で，Mikolov らにより分散表現が提案され [8]，自然言語における単語表現が連続空間のベクトルとして取り扱えるようになった．2014 年に Sutskever らにより LSTM ニューラルネットに基づく Encoder-decoder 型の機械翻訳（NMT）が提案された(図 1) [9]．連続潜在空間への埋め込み，Encoder-Decoder による Sequence- to-sequence のモデリングが機械翻訳に導入され大きな改善をもたらした．ただ，この方法では，語順の異なる言語対などで必要な過去の単語の参照が単純な再帰型ネットワークでは難しい．このことから，2015 年には Bahdanau らにより原言語，目的言語間のアライメントを効率的に表現する注視機構付き LSTM が提案された（図 2） [10]．2017 年には Encoder-decoder 型

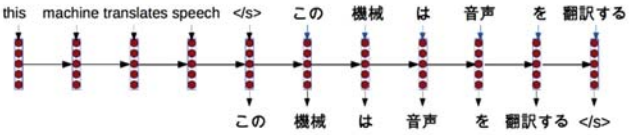

図 1　Encode-decoder 型 NMT

†1 奈良先端科学技術大学院大学　データ駆動型科学創造センター
Data Science Center, Nara Institute of Science and Technology
s-nakamura@is.naist.jp

Vol.2019-MUS-123 No.2
Vol.2019-SLP-127 No.2
2019/6/22

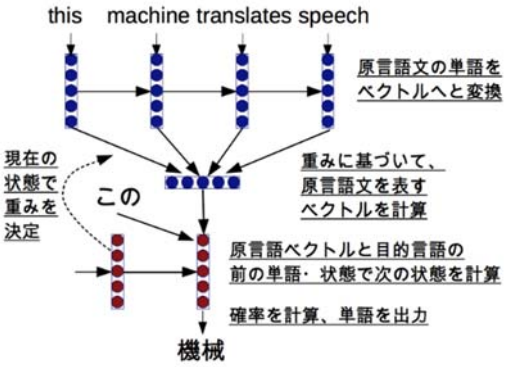

図 2 注視機構付き NMT

で問題となる計算量を，RNN を使わないことで削減する Transformer が提案された（図 3）[11]. この方法では，RNN における注視機構の代わりに自己注視(Self-attention)を使い，入力文内，出力文内のそれぞれの他の場所の情報を選択的に参照する．また，位置関係の情報を position coding により利用する．これらの工夫により Transformer は機械翻訳の性能をさらに向上させた．

3. 音声翻訳の新たな挑戦

3.1 Machine Speech Chain

著者等のグループでは，音声認識と音声合成を統合し，聴取と発声における人の処理過程である Speech Chain を End-to-end で模擬する深層学習法を試みている（図 4）[12]. この研究ではまず少量の書き起こし音声で初期の ASR と TTS を学習する．次に，書き起こしのない音声に対し，ASR+TTS により音声合成し元の音声との誤差を計算する．音声なしテキストに対しても TTS＋ASR による認識結果のテキストと元のテキストの誤差を計算する．両方の誤差を統合した後，それぞれのモデルを更新する．この方法により，現時点では話者特定ではあるが，ASR では 10k 発話の初期モデル 10%CER に対し，40k の教師なし音声，テキストデータを使用して 5%CER まで誤りが削減した．音声合成の性能についても，同時に対数ケプストラム距離が 7 から 6.2 に削減でき，ASR と TTS を統合的に学習する有効性が確認できた．さらに，音声合成に話者性を表現する潜在表現 Deep Speaker を持たせることで複数の話者の音声合成を行い，音声認識モデルのパラメータを推定する研究，さ

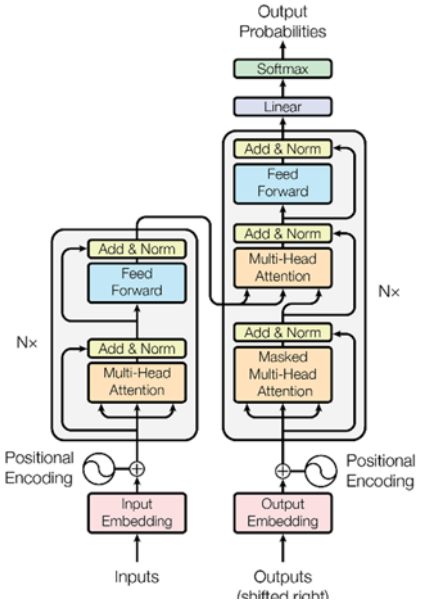

図 3 Transformer 型 NMT

らに，従来，音声合成の誤差を音声認識のパラメータ更新に利用できなかったが連結して学習が可能になる改善などを行って音声認識と音声合成を連結して相互利用することで Semi-supervised 学習ができることを示している．

3.2 End-to-end 音声翻訳

音声翻訳を Sequence-to-sequence の問題と捉えて音声入力から機械翻訳のテキスト出力までを End-to-end で学習する試みも進められている．音声入力で対象言語のテキストへの直接音声翻訳の試みは，[13] [14]で行われている．この研究の一つの大きな問題は，原言語の音声と対象言語の音声あるいはテキストのパラレルデータがないことである．このため [14]では音声合成を利用する研究も始めて行われた．これらの先行研究では，フランス語と英語のような文構造が近く，語順の変換(Reordering)の少ない言語を対象にしており，英語と日本語のように文構造が SVO, SOV と異なる言語対に対しては動作が確認されていなかった．

このような言語対を対象にする End-to-end 音声翻訳は入力から出力までが遠いので直接の学習は困難と考えられる．図 5 に著者のグループで進めている Trans-coding とカリキュラム学習に基づくシステムを示す [15]．この学習では，Phase 1 で音声認識の学習を行い，Phase2 の Fast track では，音声認識の Encoder と Attention と機械翻訳 Decoder を

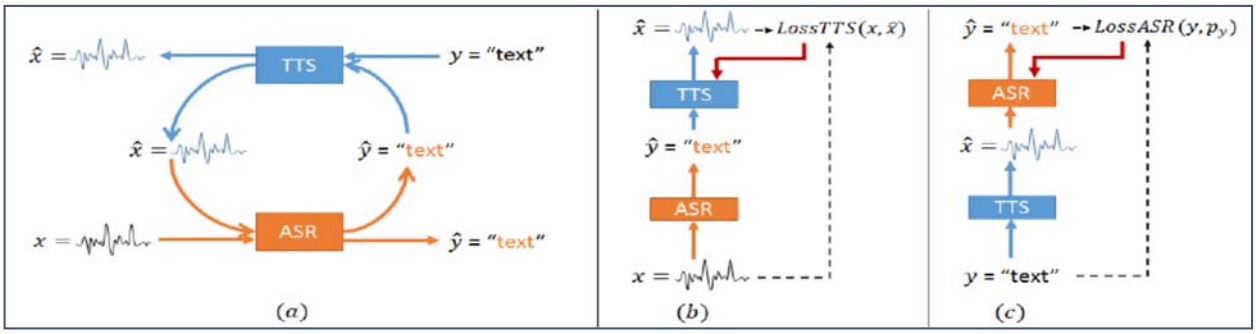

図 4 Machine Speech Chain (a)ASR-TTS 統合、(b)音声のみから TTS 学習、(c)テキストのみから ASR 学習

Vol.2019-MUS-123 No.2
Vol.2019-SLP-127 No.2
2019/6/22

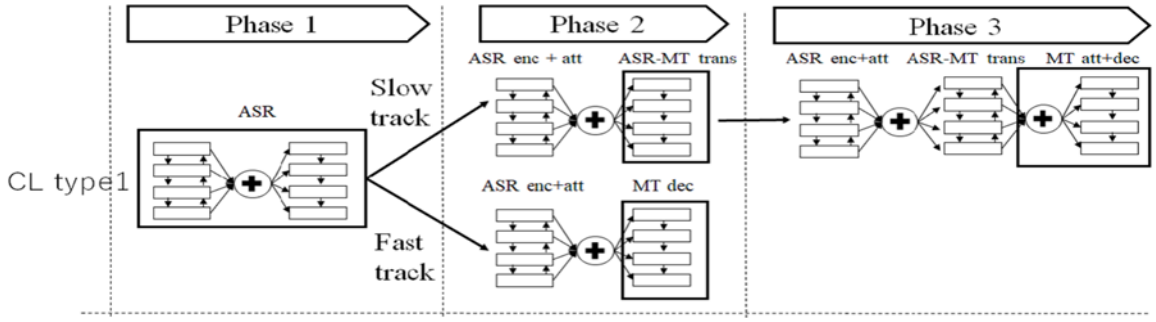

図 5　Direct Speech-to-Text Translation

組み合わせ，機械翻訳 Decoder を学習する．Slow Track の Phase2 では，音声認識の Encoder と Attention と,音声認識-機械翻訳 Transcoder を組み合わせ，Transcoder のみを学習する．最後に機械翻訳 Attention と Decoder を接続し学習する．BLEU+1 による翻訳性能評価結果を図 6 に示す．左から MT のみ, 音声認識結果入力の機械翻訳, 音声翻訳の End-to-end 学習，Fast Track，Slow Track の結果である．直接学習は非常に困難であるが, カリキュラム学習により End-to-end での学習が可能になっている．

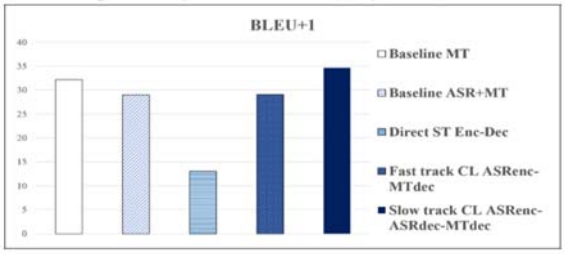

図 6　Direct 音声翻訳の性能

3.3　パラ言語情報の音声翻訳

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．筆者のグループでは, 図 7 に示すように, 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，入力発話の強調を抽出する．音声認識の結果と強調度の系列をテキスト翻訳＋強調翻訳モジュールにより変換し，目的言語で音声を合成する [16]．このテキスト翻訳＋強調翻訳モジュールには, 図 8 に示す Sequence-to-sequence の LSTM を用いる．図中，wは原言語単語列，P は品詞列，λは単語

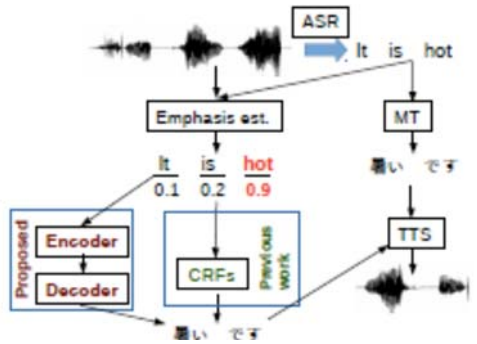

図 7　強調情報の抽出と翻訳音声への付与

の強調度合いを示す．この方法を適用した音声の主観評価

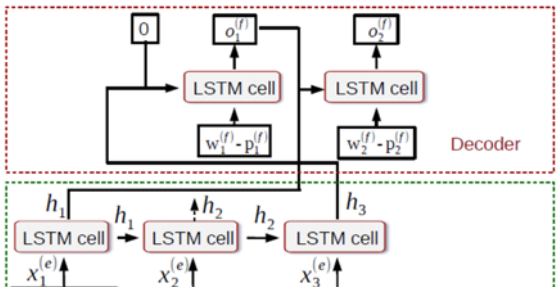

図 8　強調変換付き Encode-decoder

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．

4.　音声同時通訳へ

4.1　通訳と翻訳の差

人間による翻訳文と同時通訳文には大きな違いがある [17] [18]．翻訳文を作成する際には訳出の時間拘束がなく，前後の関係，文法を考慮した訳文が作成できる．それに較べて，同時通訳文は，聞いた音声をその場で時間遅れを最小限にし, 必要なら文の形を変形して通訳音声を出力する．次に「」から翻訳文と通訳文の文例を示す．それぞれの文を構成し翻訳の単位になるフレーズに括弧付きで番号を振っている．

(1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplies (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.

この文の翻訳文は次のようになる．

(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) 大量の難民達の (5) 世話をするための (4) 十分な食料や水, 宿泊施設, 医療品が (3) ないと (2) 言っています.

この処理を，認知負荷の度合いで調べてみることにする．図 9 に原言語のフレーズが, 記憶されてどのフレーズから訳出されていくかが示されている．下は翻訳者の記憶の必要チャンク数である．これによると必要チャンク数は 8 となる．しかしながら，人間の短期記憶，とくに，通訳のようなリアルタイムで次々と次発話の入力が入ってくる場合の短期記憶は約 3 程度と言われているため，このような翻訳は人間の通訳者には困難である．一方，実際の通訳者の

通訳文は下記のようになる．

(1) 救援担当者達の (2) 話では (4)食料，水，宿泊施設，医薬品が， (3) 足りず (6) 大量の難民達の (5) 世話ができないとのことです．(7) 難民達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．

必要な記憶チャンクについても同様に図 11 のようになる．図に示されているように，通訳者は巧みにフレーズを文として区切ったり，予測したりしながらチャンクが 3 を越えないように，通訳を実行していることがわかる．

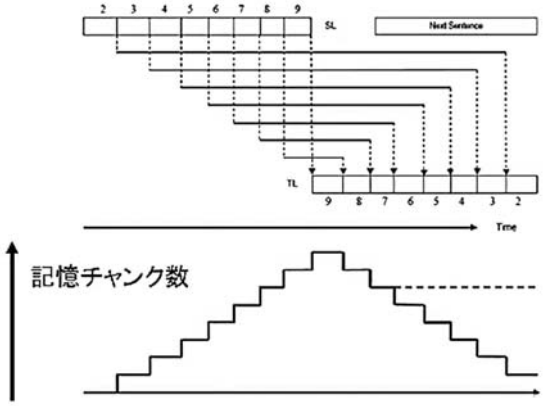

図 9　翻訳時記憶チャンク

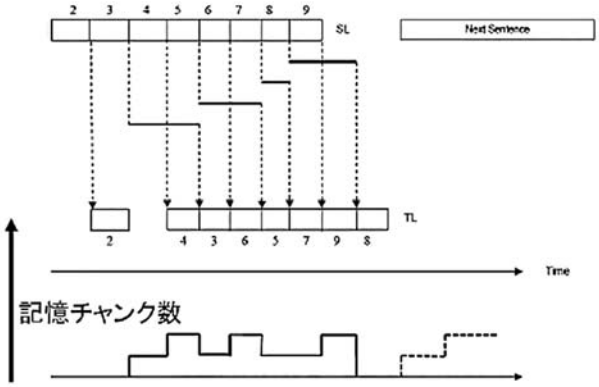

図 11　通訳時記憶チャンク

4.2 自動通訳

前節に示したように通訳と翻訳には大きな違いがある．このため，従来の機械翻訳と異なり，自動通訳システムは，①翻訳単位：文を分割して翻訳，②発話内容の予測，③表現，内容の選択　を行う必要がある．通訳においてなるべく遅延なくかつ翻訳品質を劣化させない翻訳タイミングを決定する必要がある．これまで，長いポーズなどの音響特徴に基づく手法，原言語の句読点などの言語的特徴に着目した手法, 両言語の特徴に着目した手法などが存在する．著者らのグループでは，これまで主として言語特徴に注目し 3 つの方法を検討してきた．

4.2.1 右確率を利用する方法

統計的機械翻訳では，翻訳モデルのフレーズテーブルの確率をもとに，翻訳時に原言語のフレーズチャンクが次の単語を見ずに順方向に翻訳できる場合の確率が右確率として計算できる．この右確率を用いて翻訳出力，待機を決める同時通訳手法を提案した [19]．表 1 に例を示す．右確率が高いと単語順の逆転が起らず待機，そのまま単語を翻訳

表 1　右確率を用いた同時通訳

Source	Target	RP
watashi	I	0.8
watashi ha	I	0.9
otoko	man	0.2
otoko desu	am a man	0.6
nan	what	0.9
nan ji	what time	0.7
na ji kara	from what time	0.5
pure-	play	0.2
deki	can	0.7
deki masuka	?	0.95

出力する．右確率が低いと待機し（”watashi ha otoko”の時点で右確率が 0.2），右確率が一定値を越えるまで次のフレーズを読み込み翻訳する．この方法を用いて，講演データ（TED 講演）の翻訳性能の評価（英日）を行った．図 10 に結果を示す。横軸は遅延、縦軸は翻訳性能である。結果として、経験年数 1 年のプロの同時通訳者と同等の翻訳性能であることが示された [20].

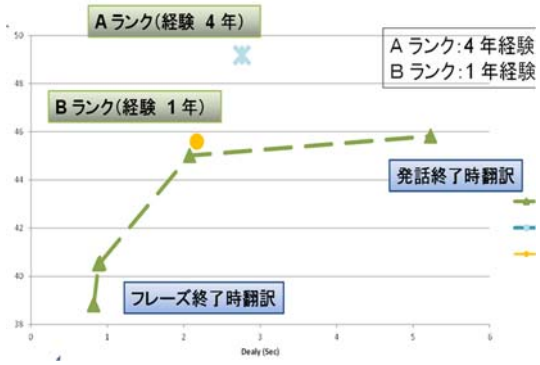

図 10 右確率法による同時通訳の遅延と性能

4.2.2 構文構造から翻訳出力タイミングを予測

この方法では，入力単語列に対し逐次構文解析を行い，次発話の部分木構造を現時点までの構文解析結果から予測し，その内容によって訳出するか，待機するかを決定し，翻訳を行う [21]。図 12 に示すように、逐次構文解析にはシフトリデュースパーザを用い，次の文要素の予測には，単語位置，チャンク内の単語列と形態素，構文木の情報を素性として SVM を用いて、訳出を行うか待機するかの予測を行っている．

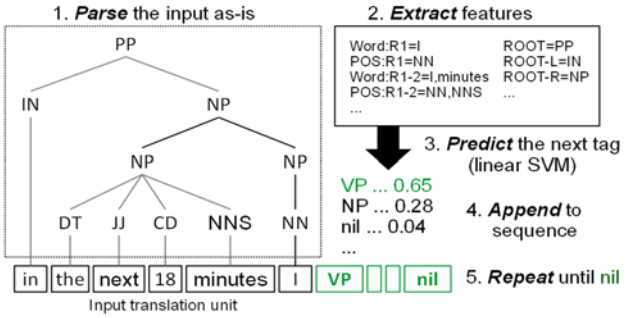

図 12　部分構文木予測による翻訳タイミング制御

4.2.3 Wait-k アルゴリズムによるニューラル同時通訳

上記の 2 つの方法は統計翻訳をベースとしたものであったが，ニューラル翻訳ベースの同時通訳法についてもいくつか検討されている．Gu et al. [22] は，既存の翻訳システムに対して 1 単語を入力する READ と 1 単語を訳出する WRITE の 2 つのアクションを翻訳タイミング決定用に設定し，強化学習によって学習する手法を提案している．また，Alinejadet al. [23] ではこの手法を拡張し，PREDICT という次に入力される単語を予測するアクションを追加した手法を提案している．これに対し，Ma et al. [24] では，“ Wait-k" モデルと呼ばれる非常にシンプルな手法が提案している．このモデルは原言語側の文の入力に対して常に k トークン遅れた状態でリアルタイムに翻訳文の生成を行う．この方法により翻訳を行う機構と動詞などの予測を行う機構の両方を統合して扱うことができ，それを Sequence-to-sequence で学習することができる．また，k を変化させることで遅延の大きさを調整することができるという利点もある．著者らは，”Wait-k" モデルを英語から日本語のような語順の異なる言語対への適用を試みた [25]．この結果を表 2 に示す．遅延トークン数が 3 または 5 で，BLEU で 10 程度の劣化で，短い語順の入れ替えに対応した翻訳出力ができることが確認できた．

表 2　Wait-k モデルによる遅延と性能

モデル	遅延トークン数 k	BLEU
Attention EncDec	(29.86)	35.70
“Wait-k” モデル	3	20.21
	5	23.01

5. 同時通訳の基盤研究

著者らは，同時通訳を含む次世代の音声翻訳へ向けた基盤研究を進めるため，音声認識，音声合成，機械翻訳，同時通夜を含む専門家を結集して科研費のプロジェクトを 2017 年から 5 年計画で実施している．この研究プロジェクトでは，講演，講義，会議を対象に，①雑音下での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，②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，③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，④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，⑤同時通訳，ビデオ翻訳コーパス構築とプロトタイプシステムを構築することを目標にしている．特に，⑤では全体で 400 時間の同時通訳コーパスの構築を目標に英日，日英の同時通訳コーパスの構築を進めている．

6. おわりに

これまでの発話ごとに翻訳を行う音声翻訳と比べ，人間の同時通訳者の行う通訳は非常に高度な知的処理である．一発話内の翻訳タイミングの問題だけでなく，過去の文脈の問題，常識，文や知識の問題，そして，コミュニケーションを考える際には，即時性，パラ言語・非言語情報，文脈，対話制御など要因の考慮が不可欠である．人間と機械の処理系は大きく異なるものの，人間の同時通訳者の振る舞いの解析，認知負荷の計測や同時通訳のメンタルモデルの研究も進めながら，自動音声同時通訳の研究を推進していく予定である．

謝辞

知能コミュニケーション研究室の教員，スタッフ，学生諸君にこの場を借りて深謝する．また，本研究は，JSPS 科研費 JP24240032，および JP17H06101 の助成を受けた．

参考文献

[1] 中村　哲編著, 音声言語の自動翻訳, コロナ社, 2018.

[2] 中村　哲, “ 話し言葉の音声翻訳技術, ” 第 96 巻 (11), pp. 865-873, 11 2013.

[3] 中村　哲, “ 音声翻訳概観, ” *電子情報通信学会誌,* pp. 702-709, 8 2015.

[4] A. Graves, et al. “ Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, ” *ICML,* 2006.

[5] A. Oden, et al., “ WaveNet: A generative model for raw audio, ” *arXiv:1609.03499,* 2016.

[6] Y. Wang, “ TACOTRON: Towards Endo-to-end Speech Synthesis, ” *arXiv:1609.03499,* 2016.

[7] J. Shen, et al., “ NATURAL TTS SYNTHESIS BY CONDITIONINGWAVENET ON MEL SPECTROGRAM PREDICTIONS, ” *IEEE ICASSP,* pp. 4779-4782, 2018.

[8] T.Mikolov, et al., “ Distributed Representations of Words and Phrases and their Compositionality, ” *NIPS,* 2013.

[9] I.Sutskever, “ Sequence to Sequence learning with Neural Networks, ” *NIPS,* 2014.

[10] D. Bahdanau, et al., “Neural Machine Translation by Jointly Learning to Align and Translate,” *ICLR,* 2015.

[11] A. Viswani, et al.,“Attention Is All You Need,” *NIPS,* pp. 5998-6008, 2017.

[12] A.Tjandra, S.Sakti, S.Nakamura, “Listening while Speaking: Speech Chain by Deep Learning,” *IEEE ASRU,* 2017.

[13] L.Duong, et al., “An Attentional Model for Speech Translation without Transcription,” *NAACL HLT,* 2016.

[14] A. Berard, et al., “Listen and Translate: A Proof of Cencept for End-to-end Speech-to-text Translation,” *CoRR,* p. vol. abs/1612.01744, 2016.

[15] T. Kano, et al., “Structured-based Curriculum Learning for End-to-end English-Japanese Speech Translation,” *INTERSPEECH,* 2017.

[16] Q.T.Do, et al., “Preserving Word-level Emphasis in Speech-to-speech Translation,” *IEEE Transactions on Audio, Speech and Language Processing,* 2017.

[17] 水野　的, 同時通訳の理論, 朝日出版社, 2015.

[18] 水野　的,“Simultaneous Interpreting and Cognitive Constraint,” 青山大学紀要　2017.

[19] T.Fujita, et al.,“Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation,” *INTERSPEECH,* 2013.

[20] H. Shimizu, et al., “Constructing a Speech Translation System using Simultaneous Interpretation Data,” *IWSLT,* 2013.

[21] Y. Oda, et al., “Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents,” *ACL,* 2015.

[22] J. Gu, et al., “Learning to translate in real-time with neural machine translation,” *EACL,* pp. 1053-1062, 2017.

[23] A. Alinejad, et al., “Prediction improves simultaneous neural machine translation,” *EMNLP,* pp. 3022-3027, 2018.

[24] M.Ma, et al., “Stacl: Simultaneous translation with integrated anticipation and controllable latency,” *arXiv preprint arXiv:1810.08398,* 2018.

[25] 帖佐克己、須藤克仁、中村哲,“英日同時通訳におけるニューラル機械翻訳の検討,” *言語処理学会大会,* 2019.