小特集──2020 年を見据えた多言語音声処理技術──

音声翻訳のあらたなパラダイム*

中 村 哲（奈良先端科学技術大学院大学）**

43.72.Kb, Ne, Ja, Bs

1. はじめに

音声翻訳は，話した音声をその場で聞き取り，目的言語に翻訳し，音声合成を行うことで異なる言語を話す相手に意図を伝える技術である。これまで長年の基礎・基盤研究，実際に話される音声言語の収録の試みが多く行われてきた。とりわけ，最近の深層学習，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ている。一方，機械翻訳や対話制御などの自然言語処理分野でも単語を連続空間のベクトルに写像することで連続表現することが可能となった。これにより，種々の自然言語処理が連続空間における処理として再定義され，多くの問題で進歩が見られている。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，音声翻訳研究をこのような音声言語処理の一つとして取り上げ，現状と今後の方向について述べる。

2. 音声翻訳のこれまで

音声翻訳については，幾つかの解説記事，書籍で紹介してきたので，歴史的経緯についての詳細はそちらを参照されたい [1–3]。本論では，これまでの技術の流れと今後の展開に関係する要素技術の流れにフォーカスする。

我が国においては，外国語の壁を崩すため 1980 年代から政府の援助を得て研究開発が進められてきた。音声認識[1]においては，発話における特徴量の揺らぎを混合ガウス分布で表し，時間方向の揺らぎを状態遷移で表した隠れマルコフモデル[2]による音響モデルと，N 単語の連接確率を確率文法として用いる N-gram 言語モデルを組み合わせる音声認識法が確立された。2000 年頃からは HMM と N-gram に加え，それらを有限状態トランスデューサ[3]として合成し，効率的に音声認識をする方法が提案され，定着した。

一方，1990 年代に，第 2 期ニューラルネットワーク[4]が登場し，再帰的ニューラルネットワーク[5]，畳み込みニューラルネットワーク[6]の原型である時間遅れニューラルネットワーク[7]，そして，長短期型リカレントニューラルネットワーク[8]等が提案された。しかし，音声認識で十分な性能を達成することができなかった。その後，ニューラルネットワークの研究，データの蓄積と公開，そして，GPU[9]による計算高速化が進み，2010 年代から急激に深層学習[10]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。

音声合成では音声の基本的単位である音素素片を連結し，単語アクセントによるイントネーションパターンとボコーダ型で音声合成する方法，可変長単位の音声素片の連結と藤崎モデルによるイントネーションパターンの合成による音声合成に発展した。更に，単語位置，前後音素コンテキストなどを情報に HMM をベースに音声合成を行う HMM 音声合成法が確立された。

一方，機械翻訳[11]では，文を形態素解析，構文解析により係り受け解析を行った後，規則を適用し

* New paradigm of speech-to-speech translation research.

** Satoshi Nakamura (Data Science Center, and Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, 630–0192) e-mail: s-nakamura@is.naist.jp

[1] ASR: Automatic Speech Recognition

[2] HMM: Hidden Markov Model

[3] WFT: Weighted Finite Transducer

[4] NN: Neural Network

[5] RNN: Recurrent NN

[6] CNN: Convolutional Neural Network

[7] TDNN: Time Delay NN

[8] LSTM: Long-short Term Memory

[9] GPU: Graphical Processing Unit

[10] DNN: Deep Neural Network

[11] MT: Machine Translation

て翻訳を行うルールベース翻訳，対訳文から同じ意味を有する文の用例を見つけ出し翻訳を行う用例翻訳が提案されたが，2000 年以降に用例そのものではなくフレーズの対応関係の確率を学習し翻訳を行うフレーズベース統計翻訳が定着した。そして，2014 年以降，再帰型 DNN である LSTM を入力単語列の符号化に用い，別の LSTM を出力単語列の復号化を行う機械翻訳法[12]が登場した。

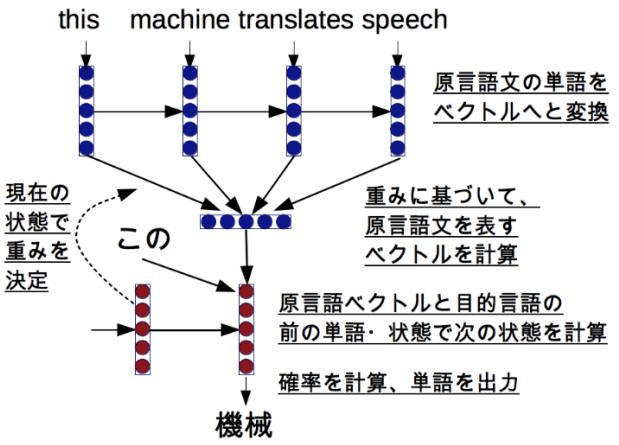
3. 音声処理，自然言語処理のいま

3.1 音 声 認 識

深層学習に基づく音声認識の最近のシステムとしては，HMM の音響モデル確率を DNN の事後確率に置き換える DNN-HMM と音声入力短区間フレームごとに音素や文字シンボル出力を生成する CTC[13]方式が主流である。いずれにしても，多層になると学習が困難であるため，CNN や LSTM を組み合わせ，出力層で統合する方法が検討されている。

Saon ら [4] は，種々の DNN の音響モデルと言語モデルを組み合わせこれまでの研究用データの性能改善を試みた。その結果，電話を通した自由会話である Switchboard や，知人同士の電話会話である Call home 自由発話タスクの単語誤り率が 5.5%, 10.3%まで改善されたと報告されている。この報告では，この性能はこれまでに知られている 5.9%, 11.3%WER のプロの人間の書き起しより高い性能だが，彼らの詳細な調査によると，正確には人間のベストの性能は 5.1%, 6.8%WER であり未だ到達していないとしている。しかし，自由発話の音声でも人間の性能に非常に近くなっていることは事実である。また，音声入力短区間フレームごとにシンボル出力を行う CTC 音声認識手法が注目されている [5]。この方法では，入力のパラメータ系列を DNN でモデリングし直接文字列を出力するように学習するシステムである。更に，エンコーダ・デコーダの構造を用いる Listen, Attend and Spell[14]が注目されている。

一方，音声合成も，Oord らにより WaveNet [6] が提案され音響モデルが波形ベースで学習できるようになった。更に Wang らは文献 [7] で，エンコーダ・デコーダの考え方で音声合成システム[15]を開発した。入力は文字列の one-hot ベクトルだが，連続ベクトルに変換された後エンコーダ・デコーダ型の注意機構付きニューラルネットワークに入力され，デコーダで振幅スペクトルを生成し，Griffin-Lim アルゴリズムで位相を生成し音声波形を合成する。この方法の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）を得たと報告されている。



図–1 エンコーダ・デコーダによる End-to-end 機械翻訳 [1]

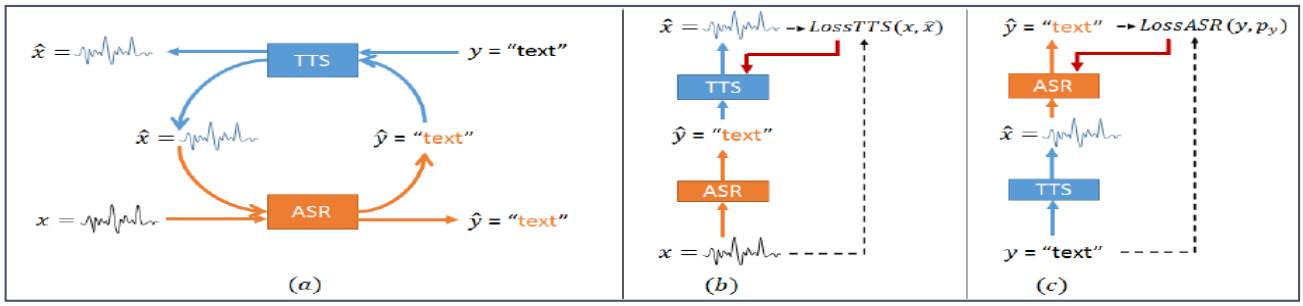
3.2 機 械 翻 訳

機械翻訳についても，2014 年頃までは統計翻訳，フレーズベースの統計翻訳，更には，構文構造を確率的に推定しながら対象言語の文字列へ翻訳する Tree-to-string や，原言語の構文木の候補から翻訳する Forest-to-string という研究が主流であった。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 らによりエンコーダ・デコーダ型ニューラルネットに基づく機械翻訳が提案された [9]。潜在空間への埋め込み，エンコーダ・デコーダ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 [9]。

この方法では，長い文の翻訳や，日英のように遠い位置の語順の入れ替えが必要な言語に問題があるため，2015 年に 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意機構付き LSTM が提案され [10]，現在の主流の方法となっている。注意機構を有するエンコーダ・デコーダの仕組みを図–1 に示す。入力単語列

[12]符号化をエンコーダ，復号化をデコーダと呼び，これらをつないだ構造をエンコーダ・デコーダと呼ぶ。この構造は Sequence-to-sequence 又は，End-to-end とも呼ばれる。
[13]CTC: Connectionist Temporal Classification
[14]LAS: Listen, Attend and Spell
[15]Tacotron と呼ばれる



図–2 End-to-end 音声認識・音声合成の模式図
(a) 全体の概要，(b) Encoder: ASR+ Decoder: TTS，(c) Encoder: TTS + Decoder: ASR

"this machine translates speech" に対し，エンコーダでこの単語列を連続ベクトル列に変換し，単語列と注意重みに基づいて原言語文を表すベクトルを求める。デコーダ側では，原言語文ベクトルと目的言語の単語履歴「この」と隠れ層の状態から次の単語の確率を求め，最も確率の大きな単語「機械」を出力する。これを繰り返して文末記号が生成されるまで目的言語単語列を生成する。

4. 音声翻訳の新たな挑戦

4.1 音声言語処理への期待

現状の音声翻訳は機械翻訳の入出力に音声認識と音声合成を統合したもので，音声認識は，性能が 100%でないため機械翻訳に誤りをもたらすモジュールとされてきた。しかし，音声認識も人間並みの聞き取り能力を達成し，音声合成も人間と変わらない音声品質を達成しつつあるいま，音声言語処理とは何かを再度考え直す時期に来ている。音声認識が正しい書き起こし結果を自然言語モジュールに送ればそれでよいのだろうか。

音声翻訳は，テキスト翻訳と異なり，書き言葉でなく話し言葉を対象にしている。書き言葉は，書く際に十分な時間があり推敲ができる。読む際にも何度も読み返して内容を理解することができる。一方，話し言葉はリアルタイムのコミュニケーションを目的にしているため，その場で意図を伝える必要があり，処理系としても実時間処理が不可欠である。この点からも，音声言語処理にはまだまだ研究の余地が残されており，再び人間の認知と関連して研究を進めていく必要がある。

4.2 Speech Chain への挑戦

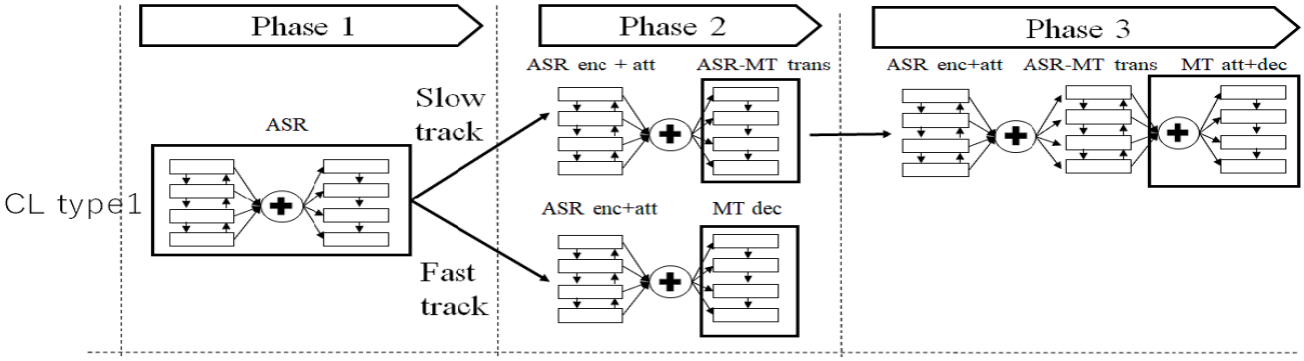
音声認識と音声合成をそれぞれエンコーダ・デコーダモデルで構成し，統合することで，脳内における Speech Chain を深層学習で模擬する研究が試みられている [11]。図–2 に処理の模式図を示す。この研究ではまず少量の書き起こし音声で初期の ASR と TTS を学習する。次に，書き起こしのない音声のみのデータに対し，ASR により単語列を求める。更に TTS により音声合成を行えば，元の音声信号と TTS 後の再生音声信号との誤差を計算することができる。逆に，音声なしテキストに対しても TTS により音声合成を行い，その音声信号を ASR により単語列を求めれば，認識結果のテキストと元のテキストの誤差（クロスエントロピー）を計算することができる。この誤差を用いて，誤り逆伝搬法により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 発話により学習した初期モデル 10%文字誤り率[16]に対し，40k の音声のみ，テキストのみデータを使用して教師なし学習を行うと 5%の文字誤り率まで誤りが削減した。音声合成の性能についても，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されている。

4.3 音声同時通訳の試み

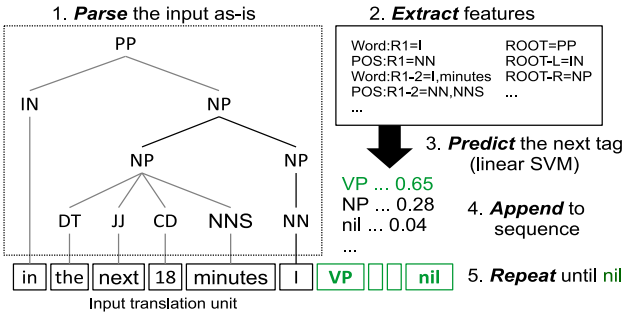
現在の音声翻訳は，発話が終了し，音声認識が終了してから機械翻訳と音声合成が行われる処理パイプラインとなっている。このため，講演のように一発話が 10 秒以上になる発話では，発話終了を検出してからでは出力が遅すぎることになる。プロの同時通訳者は発話内容を理解し，チャンキングし文構造の違いを考慮しながら適切な遅延で通訳を行っている。音声同時通訳ではこのような処理が必要となる。

日英のように文構造が違う言語対に対する同時通訳の実現にむけて，フレーズベース統計的機械

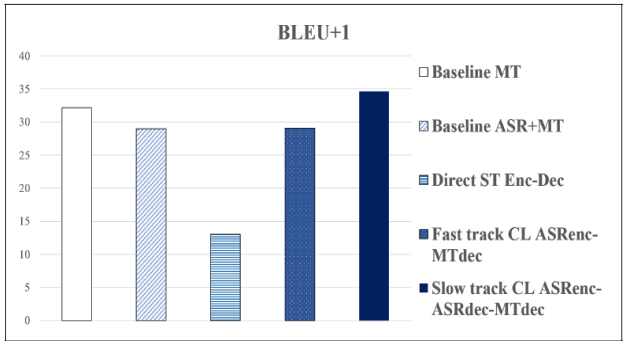
[16] CER: Character error rate



図–4　End-to-end 音声翻訳のカリキュラム学習



図–3　同時通訳のための訳出判定



図–5　End-to-end 音声翻訳の性能

翻訳における翻訳モデルの右確率[17]を用いて翻訳出力，待機を決める同時通訳手法が提案されている [12]。音声認識の結果を単語ごとに受け取ると仮定し，講演データ（TED 講演）の翻訳性能の評価（英日）を行ったところ，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 [13]。更に，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，その内容によって待機せずに訳出するか，待機するか，を SVM により決定する手法（図–3）も提案されている。部分的に構文解析済みなので，得られた部分フレーズは，Tree-to-string 翻訳モデルで翻訳される [14]。

4.4 End-to-end 音声翻訳

音声翻訳は原言語の入力音声から対象言語の音声への写像の問題と捉えることができ，エンコーダ・デコーダで全体を一つの End-to-end モデルで学習できる可能性がある。この方向の研究として，音声入力から機械翻訳のテキスト出力までをエンコーダ・デコーダモデルで学習する試みも進められている [15]。一般には音声翻訳は入力から出力までが遠いのでエンコーダ・デコーダモデルでの学習は困難である。文献 [15] ではカリキュラム学習に基づいてエンコーダ・デコーダモデルを逐次学習する（図–4）。この学習では，逐次学習していく際の方法として二つの手順（Fast Track, Slow Track）を比較している。まず，Phase 1 で音声認識の学習を行い，Fast track の Phase 2 では，音声認識の学習済みエンコーダ，注意機構と機械翻訳デコーダを組み合わせ，機械翻訳デコーダの部分を学習する。Slow Track の Phase 2 では，音声認識の学習済みエンコーダ，注意機構 と，音声認識–機械翻訳の合成を行う変換器[18]を組み合わせ変換器のみを学習する。最後に機械翻訳注意機構とデコーダを接続し再学習する。図–5 に BLEU+1 [16] による翻訳性能評価結果を示す。左から MT のみ，音声認識結果入力の機械翻訳，音声翻訳の直接 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。

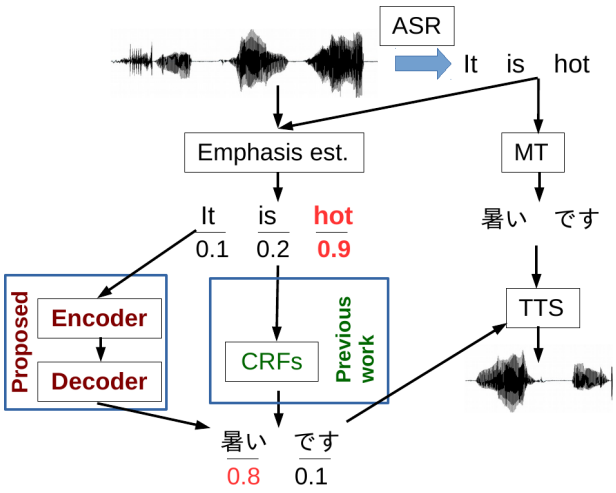
4.5 パラ言語情報の音声翻訳

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

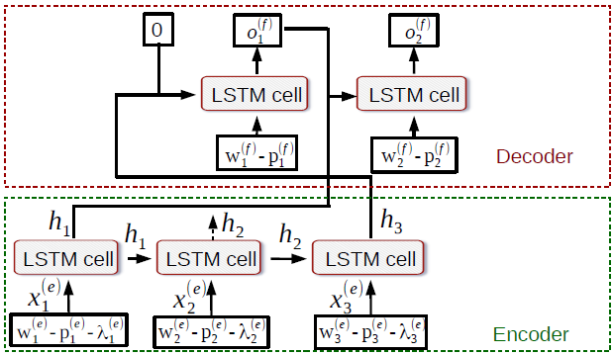
[17]語順の逆転が起こり易いかの確率を学習データで学習

[18]変換器：Transcoder

ために重要である。文献 [17] の研究では，図–6 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。文献 [18] では，更に，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。この模式図を図–7 に示す。ここで，w は原言語単語列，p は品詞列，λ は単語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示されている。



図–6 強調とテキストの強調音声翻訳



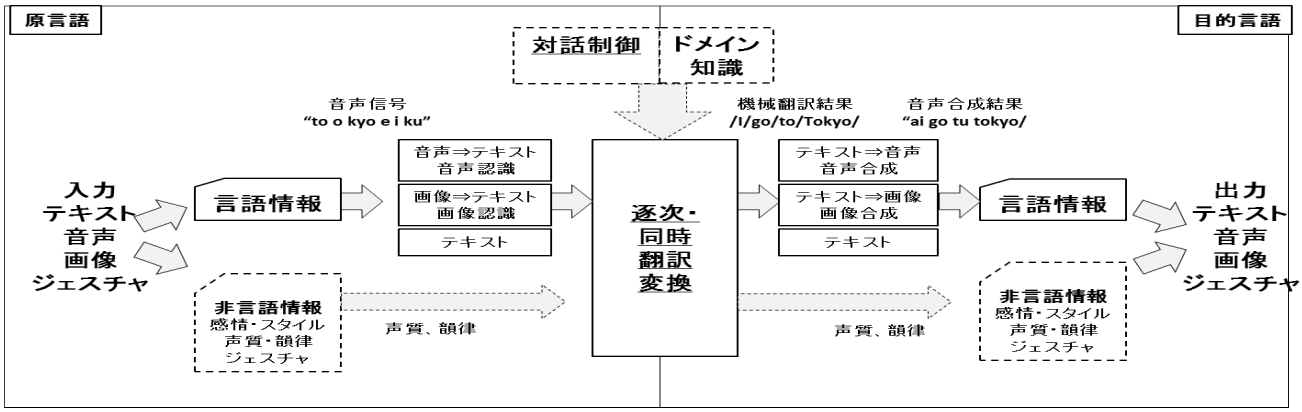
図–7 End-to-end 強調音声翻訳

5. おわりに

現在の音声翻訳が一発話ずつ独立に処理を行なうのとは異なり，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，対話制御などが不可欠である。音声翻訳を多言語対話システムとして捉えると，対話の即時性，文脈を考慮した意図，対話目標，話題，対話状態の推定，コミュニケーションを成立させるための発話者間の知識の共通基盤のモデリング，自動学習なども課題となる。図–8 に，次世代の音声翻訳システムの構成予想図を示す。発話者が表出する音声，テキスト，ジェスチャ，表情，及び，その状況を考慮して逐次・同時に，ドメイン知識，対話制御を踏まえながら，情報の翻訳，変換を行って，目的言語における言語，パラ言語情報として出力し，リアルタイムに意図を伝えるシステムとなると考えられる。このように，本当に言語の壁を越えていくためには，今後更に多くの研究が残されていると考えている。

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，及び JP17H06101 の助成を受けた。



図–8 次世代音声翻訳の構成

文　献

[1] 中村　哲 編著, 音声言語の自動翻訳, 音響サイエンスシリーズ, 日本音響学会編（コロナ社, 東京, 2018）.

[2] 中村　哲, “話し言葉の音声翻訳技術,” 信学会誌, **96**, 865–873 (2013).

[3] 中村　哲, “音声翻訳概観,” 信学会誌, **98**, 702–709 (2015).

[4] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi and P. Hall, “English conversational telephone speech recognition by humans and machines,” *arXiv:1703.02136* (2017).

[5] A. Graves, S. Fernandez, F. Gomez and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” *Proc. Int. Conf. Machine Learning 2006*, pp. 369–376 (2006).

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *arXiv:1609.03499* (2016).

[7] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, “TACOTRON: A fully END-to-END speech synthesis model,” *arXiv:1609.03499* (2016).

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Proc. Neural Inf. Process.*, *NIPS 2013* (2013).

[9] I. Sutskever, O. Vinyals and Q. V. Le, “Sequence to sequence learning with neural networks,” *Proc. Neural Inf. Process.*, *NIPS 2014* (2014).

[10] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473* (2014).

[11] A. Tjandra, S. Sakti and S. Nakamura, “Listening while speaking: Speech chain by deep learning,” *arXiv:1707.04879* (2017).

[12] T. Fujita, G. Neubig, S. Sakti, T. Toda and S. Nakamura, “Simple, lexicalized choice of translation timing for simultaneous speech translation,” *Proc. Interspeech 2013*, pp. 3487–3491 (2013).

[13] H. Shimizu, G. Neubig, S. Sakti, T. Toda and S. Nakamura, “Constructing a speech translation system using simultaneous interpretation data,” *Proc. Int. Workshop Spoken Language Translation* (2013).

[14] Y. Oda, G. Neubig, S. Sakti, T. Toda and S. Nakamura, “Syntax-based simultaneous translation through prediction of unseen syntactic constituents,” *Proc. Assoc. Comput. Linguist.*, pp. 198–207 (2015).

[15] T. Kano, S. Sakti and S. Nakamura, “Structured-based curriculum learning for End-to-end English-Japanese speech translation,” *Proc. Interspeech 2017*, pp. 2630–2634 (2017).

[16] C.-Y. Lin and F. J. Och, “ORANGE: A method for evaluating automatic evaluation metrics for machine translation,” *Proc. Coling 2004*, pp. 501–507 (2004).

[17] Q. T. Do, T. Toda, G. Neubig, S. Sakti and S. Nakamura, “Preserving word-level emphasis in speech-to-speech translation,” *IEEE Trans. Audio Speech Lang. Process.*, **25**, 544–556 (2017).

[18] Q. T. Do, S. Sakti and S. Nakamura, “Toward expressive speech translation: A unified sequence-to-sequence LSTMs approach for translating words and emphasis,” *Proc. Interspeech 2017*, pp. 2640–2644 (2017).