解説

音声認識技術の変遷と最先端 *

--深層学習による End-to-End モデル--

河原達也（京都大学）**

43.72.Ne

1. はじめに

人工知能（AI）ブームが続いている。前回約 30 年前のブームを知っている世代は、いつまで続くのか心配の向きもあるが、今回は自動運転、金融・電子商取引、医療などの広範な社会基盤に浸透しており、IoT やビッグデータと連携して「第４次産業革命」を起こしているということなので、しばらく続くのであろう。

メディアで「人工知能」と呼ばれているものを概観すると（多少ひいき目であるが）、機械学習しているものと、音声で対話しているものが多い。これら両者は、音声認識の基盤と応用となっている。著者は、音声認識自身は人工知能の範疇でないと捉えているが、音声認識は人工知能において不可欠な一つとなっている。

今の人工知能ブームの最大の推進力となっているのが、機械学習とりわけ深層学習である。実際に、深層学習は音声認識においても革新的な進歩をもたらしたし、深層学習のインパクトを示した最初の一つが音声認識である。

著者は音声認識に関する教科書[1]や解説記事[2][3]を数年おきに執筆しており、また本学会誌でも約１年半前に小特集が組まれた[4-6]が、深層学習による展開はめまぐるしい[1]。特に最近、研究の主流となっている End-to-End モデルは、著者が音声認識研究を開始した 30 年以上前から用いられている方法論・アーキテクチャを一新するところまで到っている。本稿では、これらの変遷と「最先端」のモデルについて解説する。

2. 音声認識の方法論の変遷

音声認識研究の歴史は 60 年以上に渡る。京都大学では 1960 年頃に単音節単位の認識を行う「音声タイプ」が構築されている[7]。その後、音声認識に有効な音響特徴量と、DP マッチングに代表される動的パターンのマッチング手法に関する基礎的な研究が行われた。これらは、テンプレートベースの方法であり、多数話者のバリエーションをモデル化するには不十分であった。

これに対して、確率的なモデルを導入することにより解決が図られた。DP マッチングを拡張した形で隠れマルコフモデル(HMM)が導入され、その改良が様々に行われた。まず、HMM の各状態の音響特徴量のパターンを混合正規分布(GMM)でモデル化することが導入された。そして、これを最尤推定する代わりに、識別誤りを最小化するように学習（識別学習）する方法が検討された。2000 年代に実用化された音声認識システムは、GMM-HMM の識別学習に基づくものである。一方、言語モデルについては、単語の連接規則（文法）をオートマトンで記述したものから、その遷移を確率的なものにし、その確率をコーパスから最尤推定する N-gram モデルに移行した。

その後、深層学習の導入が進められた。音声認識にニューラルネットワークを用いることは 1990 年頃から研究されていたが、主流になったのは 2010 年以降である。音響モデルについては、GMM による確率計算をディープニューラルネットワーク(DNN)に置き換えた DNN-HMM が、言語モデルについては、リカレントニューラルネットワーク(RNN)を N-gram と併用するモデルが一般的になっている。最近は、RNN を発展さ

---

* State of speech recognition technology.
– Deep learning and end-to-end modeling–

** Tatsuya Kawahara (School of Informatics, Kyoto University, Kyoto 606-8501)
e-mail: kawahara@i.kyoto-u.ac.jp

[1] 本解説記事の内容の多くは[4-6]と重なるが、[4-6]ではそもそも音響モデルと言語モデルが別々の記事になっており、両者を一体的に扱う End-to-End モデルの話はほとんどない。

表 1　音声認識の方法論の変遷

第 1 世代	1950～1960 年代	ヒューリスティック
第 2 世代	1960～1980 年代	テンプレート　(DP マッチング，オートマトン)
第 3 世代	1980～1990 年代	統計モデル　(GMM-HMM, N-gram)
3.5 世代	1990～2000 年代	統計モデルの識別学習
第 4 世代	2010 年代	ニューラルネット　(DNN-HMM, RNN)
4.5 世代	2015 年～	ニューラルネットによる End-to-End

せた LSTM (Long Short-Term Memory；4.3 節参照)のみで音声認識を行う End-to-End の枠組みが研究されている。

以上の変遷をまとめたのが表 1 である。世代の定義は古井[8]に従ったものであるが、第 4 世代以降は著者が追加したものである。

3.　音声認識の定式化とアーキテクチャ

音声認識は、音声の特徴量（の時系列）Xが与えられたときにその単語列 Wを同定する問題である[1]。すなわち、$p(W|X)$を最大化する W を求めればよい。ただし、Xも 　Wも系列である点が、画像認識などの多くのパターン認識の問題と異なる点である。なお、W については、単語列でなく文字列でもよいが、語彙の知識は認識の上できわめて有効であるし、単語がわからないと、検索・翻訳・対話などの後段の処理も意味をなさないので単語単位の系列が一般に用いられる。

3.1　ベイズ則に基づく階層モデル

従来は、この $p(W|X)$をベイズ則で書き換えることにより定式化してきた。

$$p(W|X)=\frac{p(W)p(X|W)}{p(X)} \qquad (1)$$

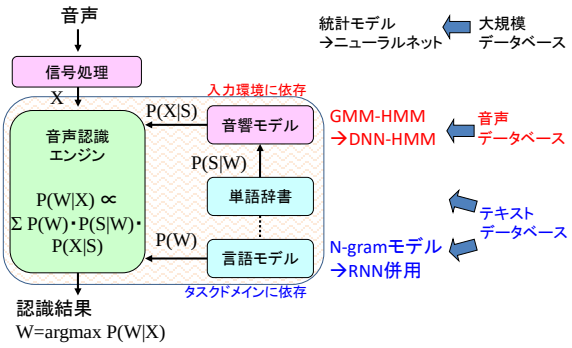
ここで Wの同定において、分母 $p(X)$は無関係なので、分子の 2 つの項のみを計算すればよい。$p(W)$は（その言語あるいは状況において）単語列 Wが生成される先験的な確率（＝言語モデル確率）であり、$p(X|W)$は単語列 Wから音声の特徴量 X が生成される確率（＝音響モデル確率）である。実際には、各単語は音素などのサブワード単位 S でモデル化され、単語と音素の関係は辞書で決定的に与えられる($p(S|W)=\{1,0\}$)ので、以下のようになる。

$$p(W)p(X|W)=\sum_S p(W)p(S|W)p(X|S)\approx \max p(W)p(X|S) \qquad (2)$$

この定式化は、1990 年頃から標準的になり、その後四半世紀以上にわたって、世界中の言語において普遍的に用いられ、あらゆる教科書の冒頭に記述されてきた。この理由として、$p(W|X)$の確率分布を直接推定することが難しいこと、及びパターン認識処理で実現される音響モデルと自然言語処理で実現される言語モデルに分離した方が問題を扱いやすいことが挙げられる。

これに基づく音声認識システムのアーキテクチャを図 1 に示す。各モデルが階層的に構成され、各モデルで計算される確率を順次結合することによって $p(W|X)$が計算される。これを本稿では階層モデルと呼ぶ。ここでは、各モデルは生成モデルとして定式化・学習される。

図1 従来の音声認識システムの構成



3.2　End-to-End モデル

これに対して図 2 に示すように、上記のモジュールを統合的にニューラルネットワークで構成して、$p(W|X)$あるいは $p(S|X)$を直接推定する End-to-End の枠組みが研究されている。図 2 では、音響モデルを音素状態の認識部(DNN)と音素などのモデル(HMM)にさらに分けている。

図2 End-to-End音声認識

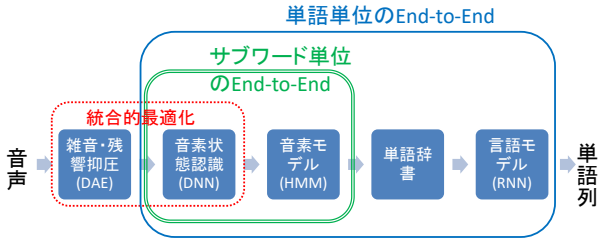
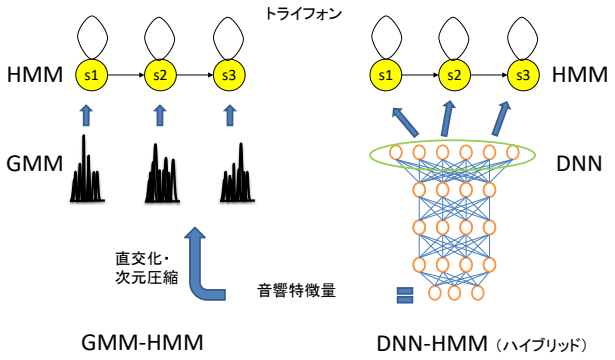


図3 GMM-HMMとDNN-HMM



この枠組みで最も早く実現されたのが、音素や文字などのサブワードを出力の単位とするLSTM を用いて、その出力系列を縮約する CTC (Connectionist Temporal Classification)である[11] (5.1 節参照)。これとは別に、LSTM で入力系列をいったん符号化した後に、サブワード系列に復号化する注意機構モデル[12] (5.2 節参照)も導入されている。これらのモデルはサブワード単位の言語モデルを暗黙的に包含し、$p(S|X)$を直接計算するものと捉えられる(図 2 の二重線枠部分；5 章で詳述)が、単語辞書や言語モデルはこの後に適用する必要がある。

さらに最近では、これらのモデルを発展させて、単語を出力の単位としたモデル化も検討されている。これにより、$p(W|X)$を直接計算することができる(図 2 の実線枠部分；6 章で詳述)。

これとは別に、雑音や残響抑圧のためのニューラルネットワーク(デノイジングオートエンコーダ(DAE)などで構成)と音素状態認識の DNN を統合して学習することも研究されている(図 2 の点線枠部分)。

本稿ではフロントエンドは扱わず、音素状態認識以降の処理(図 2 の実線枠内)について述べる。

4. 階層モデルにおける深層学習の導入

深層学習の音声認識への導入は当初、階層モデルの枠組みで実現された。最も直接的な方法は、HMM における各音素状態の GMM による確率計算を DNN に置き換えるもので、DNN-HMM ハイブリッドモデルと呼ばれる[9,10]。

4.1 DNN-HMM ハイブリッド音響モデル

DNN-HMM の構成を図 3 に示す。DNN の入力は音響特徴量であり、当該フレームの前後のフレームの周波数特徴量も結合して与える。出力層のノードは、HMM の各状態に対応付けられるが、一般的な音声認識ではトライフォンモデルの共有状態となる。これは、先行音素と後続音素の文脈を考慮したもので、クラスタリングを行っても数千個のオーダになる。音声認識で必要となるのは、式(2)の $p(X|S)$であるが、DNN で計算される出力確率は通常 Softmax 関数を経た事後確率 $p(S|X)$の形となるので、事前確率 $p(S)$で除して、HMM に渡す。

DNN-HMM の学習手順は以下の通りである。

(1) GMM-HMM を学習する。これは、トライフォン状態のクラスタリングを含む。
(2) 学習データベースをトライフォン HMM の状態でアライメントする。
(3) アライメントされたデータ(音響特徴量と HMM の状態ラベルの対)を用いて、DNN を学習する。

HMM の状態遷移確率は(1)で推定され、各状態の事前確率は(2)で推定されるものを用い、出力確率を DNN で計算する。DNN の学習を行うための音素状態ラベルは、GMM-HMM によって作成されていることに留意する。

各種の音声認識タスクにおいて、DNN-HMM が GMM-HMM を大きくしのぐ(誤り率で概ね 20～30%の削減)ことが示され、普遍的に用いられるようになった。その後も、CNN や ResNet、さらに LSTM などを多層に積み重ねるなどの改良が行われている。

4.2 系列識別学習

上記の DNN の学習は、音響特徴量と音素状態ラベルの対が与えられて行われ、クロスエントロピー基準で行われる。これは静的なパターン認識と同じである。

一方、音声認識では、パターンの時系列に対する尤度に基づいて文認識結果が得られ、これに対して音素誤りや単語誤りが評価される。そこで、クロスエントロピー基準によりフレーム正解精度を最適化する代わりに、発話全体の認識結果系列としての誤りを最小化する学習も定式化でき、これを系列識別学習と呼ぶ。そのために、下式のベイズリスク（BR）定義する。

$$BR(X,W) = \sum_{W'} p(W'|X) \cdot loss(W',W) \qquad (3)$$

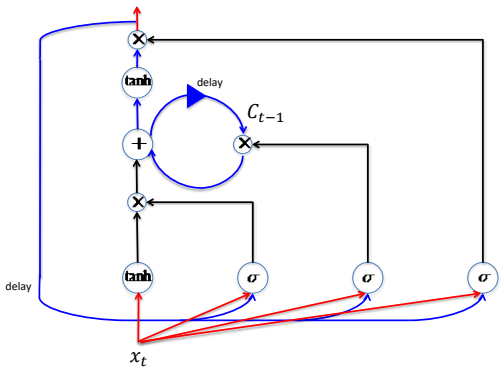
ここで $loss(W)$は損失関数で、音素誤り率などで定義されるが、DNN の学習で最も一般的に用いられているのが、HMM 状態誤りの期待値を最小化するもの(sMBR)である。上記の計算には、多数の文候補 Wを生成する必要があるが、単語グラフを生成することで効率的に実現できる。

系列識別学習の詳細な定式化と説明は文献[10]の 8 章参照。

4.3 RNN 言語モデル

音響モデルだけでなく、言語モデルにおいてもニューラルネットワークの導入が進められており、RNN が一般に用いられている。ただし、入力単語を少ないノードの数値データ（分散表現）に射影する層を別途用意する。中間層はこれと履歴を符号化したものと捉えられ、N-gram モデルと比べて非常に長い履歴を考慮することができる。ただし、N-gram モデルの方が低頻度語のスムージングが効果的に行えることもあり、N-gram モデルと併用・線形補間する場合が多い。

RNN を発展させたものとして、種々のゲートを導入した LSTM (Long Short-Term Memory) がある。これを図 4 に示す。LSTM では、内部メモリで 1 つ前の状態（分散表現）を記憶しつつ、入力・記憶の利用・出力の各過程においてゲートを設定して、情報の制御を行っている。このゲートの値（2 値に近い）も入力と 1 つ前の出力を元に決まるように学習される。LSTM は次節以降に述べる End-to-End モデルの基盤となっている。



5. End-to-End モデルによる系列写像学習

階層モデルのように多数の処理を経るのではなく、入力から出力への写像を直接（統合的に）学習するものを End-to-End モデルと呼ぶ。特に、音声認識では、入力が音声の時間フレーム長 T で規定される音響特徴量の系列、出力が音素・文字・単語などの記号列であり [2]、このように長さの異なる時系列間(seq2seq)の写像を行うものを End-to-End モデルと呼ぶ。音声合成（単語列→音響特徴量系列）や機械翻訳（単語列→単語列）でも同様の定式化が行える。

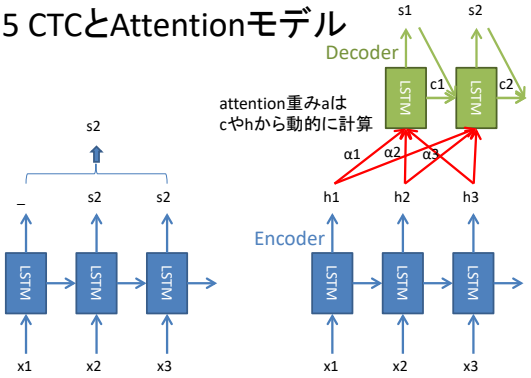
音声認識における End-to-End モデルは、HMM を一切用いずに実現できる。その方法として、現在主に採用されているのが以下の 2 つである。両者の概念を図 5 に示す。

5.1 Connectionist Temporal Classification (CTC)

HMM を用いることなく、ニューラルネットワークのみで時系列パターンを分類しようと定式化されたのが CTC である[11]。CTC では通常、音素単位や文字単位の LSTM が用いられる。ただし、どの音素・文字でもないブランク記号（_）を導入し、各音素・文字の間に挿入する。入力時間フレーム毎にこれらの記号が出力され、これを集積する。この際に、ブランク記号を消去し、時間的に連続した同一の出力記号を 1 つに縮約する操作を行う（図 5 左）。

[2] 「真の End-to-End」は音声波形を入力とすべきとする向きもあるが、本稿では音響特徴量系列を入力 Xとする。ヒトの蝸牛でも周波数分析を行っている一方で、我々が通常扱う音声波形はマイクロフォンで入力されたものである。

図5 CTCとAttentionモデル



これを確率的に定式化すると、各フレームの音響特徴量 x_t に対して各記号 s_t の事後確率を計算し、同じ記号列 S に縮約されるものの総和を計算することになる。

$$p(S \mid X) = \sum_{\vartheta \to S} p(\vartheta \mid X) = \sum_{\vartheta \to S} \prod_{t=1}^{T} p(s_t \mid x_t) \qquad (4)$$

例えば、以下の記号系列はすべて hai を表現するものとしてまとめられる。

_h_____a_____i__
_hh____aa____ii_
_h____aaaa__iii_

モデル学習の際には、正解記号系列 S（上記の例では hai）に縮約されるすべての系列の尤度の総和を求める。この尤度計算は、HMM の尤度計算と同様に、前向き・後向きアルゴリズムで効率的に実現できる。この対数尤度を基に、LSTM の各パラメータを更新する学習則が導出される。

なお、CTC では LSTM により時間フレーム間の依存性はモデル化されるが、式(4)においては記号間の関係は独立に扱われている。

5.2 注意機構(Attention)モデル

注意機構モデルは、正確には注意機構付きエンコーダーデコーダ(encoder-decoder)モデルであり、符号部と復号部から構成される(図5右)[12]。

エンコーダでは、入力フレーム毎に音響特徴量を LSTM により別の数値ベクトル（分散表現）h_t に符号化する。デコーダは、この符号化された分散表現 h_t の系列（長さ T）を入力として、音素・文字などの記号 s_l の系列（長さ L）を順次予測する。その際に、例えば最初の方の音素は音声の最初の方の情報を用いるのが有用であるので、重み α_t を付ける。これが注意機構であり、この重み自体も動的に計算され、重みを計算するパラメータは統合的に学習される。c_l は LSTM の内部メモリ状態であり、以下のように更新される。

$$\begin{aligned} c_l &= LSTM(c_{l-1}, g_l, s_{l-1}) \\ g_l &= \sum_t \alpha_{l,t} \cdot h_t \\ \alpha_{l,t} &= \exp(e_{l,t}) / \sum_{t'} \exp(e_{l,t'}) \\ e_{l,t} &= f(c_{l-1}, h_t, \alpha_{l-1}) \end{aligned} \qquad (5)$$

その上で、記号 s_l は内部状態 c_l（と g_l）に基づいて（one-hot ベクトルとして）計算される。デコーダは、文頭(sos)記号から予測を開始し、文末(eos)記号が出力されると終了する。デコーダ LSTM は次の記号を予測する際に、直前の状態 c_{l-1} に加えて、直前の記号 s_{l-1} を用いており、言語モデルの機能を包含している。

エンコーダ・デコーダ及び注意機構の学習は、正解記号系列と予測記号系列のクロスエントロピーに基づいて統合的に行われる。

以上をまとめると、入力音響特徴量系列 X をいったん分散表現の系列 H に変換した上で、$p(S|H)=p(S|X)$が最大となる記号系列 S を出力するモデルであるので、H を音素状態の尤度に代わるものと捉えると、エンコーダは音響モデル、デコーダが言語モデルに対応すると考えられる。

また、注意機構は音声と記号の対応付けを行うものと捉えられる。このアライメントを正確に行うために、順方向の LSTM のみでなく、発話末から逆方向の LSTM を併用した双方向 LSTM を用いるのが一般的であり、エンコーダではこれを多層用いることが多い。これは、CTC においても同様である。

6. 単語単位の注意機構モデル（＝オールインワンシステム）

End-to-End 音声認識に関する研究の大半は、音素や文字を単位として行われていた。語彙の制約や単語単位の言語モデルは、後処理として適用する必要がある。これは、式(2)の $p(X|S)$ を $p(S|X)$ で代用し、$p(W)$ を組み合わせる過程に相当する。

認識精度の点において、End-to-End モデルは従来の DNN-HMM ハイブリッドシステムを上回る結果はあまり得られていないが、単語辞書や言語モデルをデコーディングの早い段階で適用していないことが大きな原因と考えられる。

これに対して、1～2 年前から単語を単位とした End-to-End モデルの研究も行われている[13]。これは、語彙や言語モデルをすべて包含して、一気に $p(W|X)$を求めるもので、ニューラルネットワークのみで音声認識処理が完結し、複雑なプログラムである音声認識デコーダを一切必要としない。3.1 節で述べた従来の標準的な枠組み・アーキテクチャを一新するものである。

ただし、音素や文字に比べて、単語のエントリ数は圧倒的に多く、また出現頻度の偏りも大きいため、学習が容易でない。実際にかなり大規模な音声データベースがないと構築が困難である。また、サブワード単位の認識と比較して、音声データベース中に出現しない未知語を後で追加できないという実用的に重要な問題もある。

これに対して著者らは、文字単位のモデルと併用することで解決を図っている[14]。単語単位のモデルについては、5 節で述べたように注意機構モデルの方が言語モデルを直接的に包含するので望ましい。この注意機構モデルの正則化の効果を期待して、文字単位のモデルは CTC を採用し、エンコーダ部分を共有する構成とした。その上で、両者のマルチタスク学習を行う。また、単語単位のモデルで未知語を検出した際には、文字単位のモデルの認識結果を用いて復元する。

『日本語話し言葉コーパス』(CSJ)を用いて、約 2 万語彙のシステムを実装・評価したところ、学会講演・模擬講演ともに、DNN-HMM ハイブリッドモデルを認識率で上回ることができた。また、認識に要する時間の実時間比は 0.035 と DNN-HMM に比べて 25 分の 1 以下となった。発話終了後でないと処理を開始できないが、この処理速度であれば問題ないと考えられる。

7. 「人間と同等」の音声認識

音声認識の近年の話題として、IBM と Microsoft が電話会話音声認識のタスクで約 95%の認識率を達成し、「人間と同等の認識精度を実現した」と発表したことが挙げられる。「人間と同等」の能力については様々な議論があると思われるが、これらのシステムはかなり複雑な階層モデルに基づくものである。

これに対して、最後に述べた単語単位の End-to-End システムはきわめて単純で、単語を知覚の単位としている人間に近いともいえる。音声認識は第 2 世代の DP マッチングでは単語を単位として構成されていたが、その後汎用性と統計的学習の観点からサブワード単位に移行し、再び単語単位のモデルが検討されているのは興味深い。また、音声合成や機械翻訳なども同様のモデルで実現されつつあるのもきわめて興味深い。ただし、End-to-End モデルは音声データベースのみで学習するので、テキストデータや単語辞書などの外部の言語資源の活用が課題である。音声のみで言語を学習するのは文字を発明する以前の人間に戻った感もする。

人工知能研究には、人間の知能そのものをもつ機械を作ろうとする立場と、人間が知能を使ってすることを機械にさせようとする立場がある[15]。音声認識を含む大半の研究は後者の立場であるが、「人間と同等」に近づいた今、前者の立場について考えるのも興味深い。

文献

[1] 河原達也 編著. 音声認識システム （改訂 2 版）. オーム社, 2016.

[2] 河原達也. 音声認識技術の現状と将来展望. 電気学会誌, Vol.133, No.6, pp.364--367, 2013.

[3] 河原達也. 音声認識技術. 電子情報通信学会誌, Vol.98, No.8, pp.710--717, 2015.

[4] 篠田浩一. 音声言語処理における深層学習：総説. 日本音響学会誌, Vol.73, No.1, p. 25-30, 2017.

[5] 神田直之. 音声認識における深層学習に基づく音響モデル. 日本音響学会誌, Vol.73, No.1, p. 31-38, 2017.

[6] 増村亮. 深層学習に基づく言語モデルと音声言語理解. 日本音響学会誌, Vol.73, No.1, p. 39-46, 2017.

[7] T.Sakai and S.Doshita. The Phonetic Typewriter. Proc. IFIP Congress 62, pp.445-450, 1962.

[8] S.Furui. Selected Topics from 40 Years of Research on Speech and Speaker Recognition. Proc. INTERSPEECH, pp.1-8, 2009.

[9] G.Hinton, L.Deng, Y.Dong, G.E.Dahl, A.Mohamed, N.Jaitly, A.Senior, V.Vanhoucke, P.Nguyen, T.N.Sainath and B.Kingsbury. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, Vol.29, No.6, pp. 82-97, 2012.

[10] D.Yu and L.Deng. Automatic Speech Recognition – A Deep Learning Approach. Springer, 2015.

[11] A.Graves and N.Jaitly. Towards End-to-End speech recognition with recurrent neural networks. Proc. ICML, 2014.

[12] J.Chorowski, D.Bahdanau, D.Serdyuk, K.Cho, and Y.Bengio. Attention-based models for speech recognition. Proc. NIPS, 2015.

[13] H.Soltau, H.Liao, and H.Sak. Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition. Proc. INTERSPEECH, pp.3707-3711, 2017.

[14] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In Proc. IEEE-ICASSP, 2018.

[15] 人工知能学会. What's AI. https://www.ai-gakkai.or.jp/whatsai/