2-2-4　音声認識技術
2-2-4　Speech Recognition Technology

藤本 雅清
FUJIMOTO Masakiyo

音声認識技術の研究開発は古典的なパターンマッチング問題から始まり、機械学習技術の発展に伴い統計的音声認識が登場した。その後、深層学習技術の台頭によりハイブリッド型音声認識の研究が行われ、さらに End-to-End 音声認識に発展した。NICT においてもこのような世界的な動向に追随して研究開発を推進しており、特に多言語の音声認識を展開している。また、最新技術である End-to-End 音声認識の研究開発にも注力している。研究成果は、モバイルデバイス向けの音声翻訳アプリ VoiceTra®等、様々な場面で活用されている。本稿では音声認識のこれまでの技術発展について俯瞰し、それに伴う NICT の研究開発について述べる。

Research and development (R&D) of speech recognition technology began with solving the traditional pattern matching problem. As machine learning technology developed, statistical speech recognition methods emerged. Then, the appearance of deep learning technology led to research on hybrid-style speech recognition and has further developed into End-to-End speech recognition. NICT has been on the same boat as these worldwide trends promoting R&D of speech recognition, especially focused on ``multilingual'' speech recognition. Furthermore, R&D of state-of-the-art End-to-End speech recognition is also being conducted. Research results are used in a wide variety of products and services deployed by various organizations, including VoiceTra®, a multilingual speech-to-speech translation app for mobile devices. This paper describes an overview of the technological development of speech recognition and R&D carried out at NICT.

1　まえがき

音声認識技術は文字通り、入力された音声を文字（テキスト）に変換する技術であり、これまで連綿と研究開発が続けられてきた。近年における幾つかの技術革新により音声認識技術は急速な発展を遂げ、スマートフォン、タブレット等のモバイルデバイスにおける音声検索、音声翻訳や、スマートスピーカーでの音声操作、映像メディア等への自動字幕付与等、様々な場面で活用され我々の生活に浸透しつつある。

音声認識技術は古くは種々の Dynamic programming もしくは Dynamic time warping に基づく単純なパターンマッチング手法 [1] に始まり、1990 年代から 2000 年代前半には機械学習技術 [2] の発展に伴い、隠れマルコフモデル（HMM: Hidden Markov Model）[3] や N-gram 言語モデル [4][5]、有限状態トランスデューサー（WFST: Weighted Finite State Transducer）[6][7] 等の方法を用いた統計的音声認識 [8] が登場し、様々な技術提案がなされた。その後、2000 年代後半ごろの深層学習（Deep Learning）[9]–[11] の登場により大幅な性能改善が示された。2010 年代には深層学習に基づく音声認識 [12]–[20] の研究開発が極めて活発化し、その中で数多くの革新的な技術が提案され、音声認識が様々な形で実用化されるに至った。

このような研究開発の歴史において、NICT においても第一線での研究開発を継続的に実施して着実に技術改善を示しており、直近では最新技術である End-to-End 音声認識の研究開発に注力している。また、NICT では多言語音声翻訳技術の社会実装を至上命題としており、その成果の一つとしてモバイル端末における多言語音声翻訳アプリ VoiceTra® [21] を開発し、社会実装における実証実験の名目で無償公開している。VoiceTra® では様々な言語の音声が入力されるので、単言語ではなく多言語の音声認識が必要となる。そのため、NICT ではアジア言語を中心とした 20 言語前後の音声認識を開発し、実装している。このように NICT では第一線での研究成果の展開のみならず、様々な言語のユーザーが音声認識を利用できるよう幅

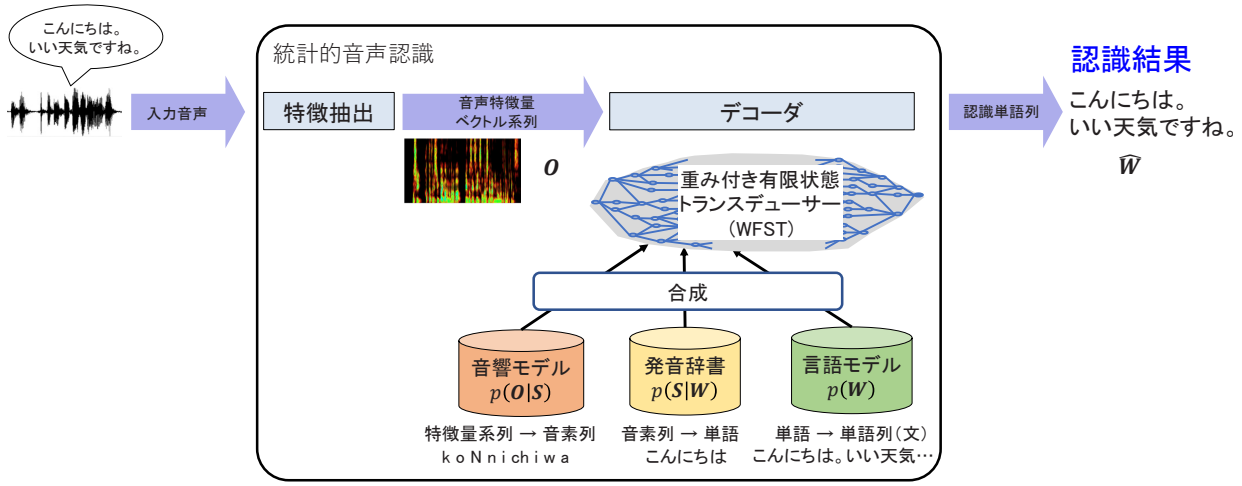


図 1　統計的音声認識の概要

広く研究開発を実施している。

本稿では上記に示した音声認識技術の研究開発の発展を俯瞰的に述べ、技術革新の要因となった幾つかの主要技術について解説する。また、このような世界的な研究開発の状況下における、NICT の取組についても紹介する。

2　統計的音声認識

まず、深層学習登場以前に主流であった統計的音声認識の概要について述べる。図 1 は統計的音声認識の概要図であり、主に特徴抽出器、デコーダ（復号器）、WFST による音声認識モデル（音響モデル、発音辞書、言語モデルの混合一体化モデル）から構成される。

図 1 においてまず、特徴抽出器は入力された 1 次元の音声信号系列を分析し、対数メル周波数スペクトルやメル周波数ケプストラム [22] と呼ばれる数十次元の特徴量ベクトル系列 $\boldsymbol{O}=\{\boldsymbol{o}_0,\cdots,\boldsymbol{o}_t,\cdots,\boldsymbol{o}_T\}$ を抽出する（Tは総入力フレーム数）。特徴量の抽出は通常 10 ms 程度の短時間フレームごとに実施する。特徴量ベクトル系列 $\boldsymbol{O}$が与えられると、デコーダは式 (1) に基づき、最尤（さいゆう）となる単語列 $\widehat{\boldsymbol{W}}=\{w_0,\cdots,w_n,\cdots,w_N\}$ を探索して出力する（Nは総出力単語数）。これにより得られた単語列 $\widehat{\boldsymbol{W}}$が音声認識結果となる [8]。

$$\widehat{\boldsymbol{W}}=\arg\max_{\boldsymbol{W}} p(\boldsymbol{W}|\boldsymbol{O}) \qquad (1)$$

式 (1) において $p(\boldsymbol{W}|\boldsymbol{O})$ は音声認識モデルに相当し、統計的音声認識では機械学習技術により大量の音声データを用いて精密な $p(\boldsymbol{W}|\boldsymbol{O})$ のパラメータ群を推定することが重要となる。

式 (1) の右辺は、音素 s（音韻を弁別する上での最小単位）の系列である音素列 $\boldsymbol{S}$ という中間表現を導入し、

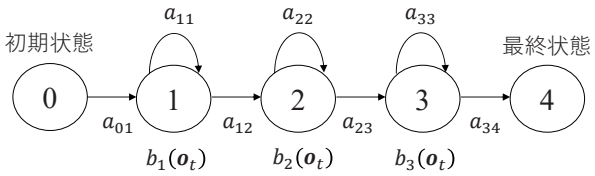


図 2　3 状態 Left-to-Right HMM。a_{ij}は状態遷移確率、$b_i(o_t)$は状態出力確率を示す。

さらにベイズの定理等を用いることにより次式の様に近似的に表現することができる。

$$\widehat{\boldsymbol{W}}=\arg\max_{\boldsymbol{W}}\sum_{\boldsymbol{S}} p(\boldsymbol{O}|\boldsymbol{S})\,p(\boldsymbol{S}|\boldsymbol{W})\,p(\boldsymbol{W})$$
$$\approx \arg\max_{\boldsymbol{W}}\max_{\boldsymbol{S}} p(\boldsymbol{O}|\boldsymbol{S})\,p(\boldsymbol{S}|\boldsymbol{W})\,p(\boldsymbol{W}) \qquad (2)$$

式 (2) のように音声認識モデル $p(\boldsymbol{W}|\boldsymbol{O})$ は 3 つの要素に分解することができ、$p(\boldsymbol{O}|\boldsymbol{S})$、$p(\boldsymbol{S}|\boldsymbol{W})$、$p(\boldsymbol{W})$ はそれぞれ音響モデル、発音辞書、言語モデルと呼ばれる。ここで発音辞書は、音素列 $\boldsymbol{S}$から単語 w を生成する確率モデルであるが一般的には確率モデルを用いず、ある一定の規則に沿った生成が行われる。以降では音響モデル、言語モデルの概要と、WFST によるモデル一体化について述べる。

2.1　GMM-HMM に基づく音響モデル

音響モデルは特徴量ベクトル系列 $\boldsymbol{O}$から最尤の音素列 $\widehat{\boldsymbol{S}}$を得るモデルである。統計的音声認識における音響モデルとして、主に HMM という有限の状態を有する非決定性の状態遷移モデルが用いられる [3]。特に図 2 に示すような 3 ～ 5 の状態を有する Left-to-Right 型の HMM が用いられることが多く、音素ごとに HMM を学習して連結することにより、最尤の音素列 $\widehat{\boldsymbol{S}}$を出力する。ここで、音素における状態とは、ある

音素sが発音された際の時間的な変化、すなわち音の立ち上がり、立ち下がり（過渡状態）、安定（定常）状態等を表現することが多い。

各状態における出力確率$b_i(\boldsymbol{o}_t)$を得るための確率分布には正規分布を適用することが多く、特に複数の正規分布を重み付け加算することにより表現された混合正規分布（GMM: Gaussian Mixture Model）[2][3]が用いられる。このように、HMMによる状態遷移構造を有し、GMMによる出力確率分布を有する音響モデルをGMM-HMM音響モデルと呼ぶ。

GMM-HMM音響モデルの学習にはBaum-Welchアルゴリズム[23][24]、音素（HMM状態）列の探索にはViterbiアルゴリズム[25][26]が用いられており、各々の詳細については文献を参照されたい。また、音素の音響的な特徴は前後に連結される音素によってその特徴が大きく異なる場合がある（調音結合[19][22]の影響）。この特徴を詳細にモデル化するため、前後の音素の影響を考慮したTri-phoneモデル[19][22]が広く利用されている。なお、前後の音素の影響を考慮しない場合はMono-phoneモデルと呼ばれる。Tri-phoneモデルを効率的に学習するため、Tree-based clustering、State tying等の様々な手法[27]–[29]が用いられており、これらについても詳細は文献を参照されたい。

2.2　N-gramに基づく言語モデル

言語モデルは、発音辞書により得られた単語$\boldsymbol{w}$から最尤の単語列$\widehat{\boldsymbol{W}}$を得るモデルである。すなわち、言語モデルは与えられた任意の記号列に対して、その言語らしさを規定するモデルとなる。音声認識においては音響モデル、発音辞書により得られる出力単語候補の言語的な妥当性を考慮することで、より高い精度での単語列$\boldsymbol{W}$（音声認識結果）の出力を可能とする。

ある単語列$\boldsymbol{W}$が生成される確率は、

$$\begin{aligned} p(\boldsymbol{W}) &= p(w_0)p(w_1|w_0)\cdots p(w_N|w_0,\cdots,w_{N-1}) \\ &= p(w_0)\prod_{n=1}^{N} p(w_n|w_0,\cdots,w_{n-1}) \end{aligned} \quad (3)$$

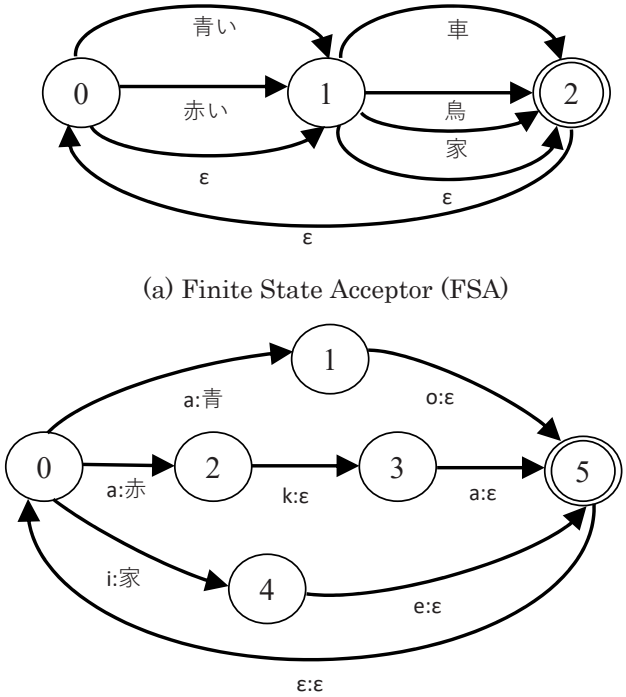
により得られるが、様々な長さの単語列、単語の組合せに対して条件付き確率$p(w_n|w_0,\cdots,w_{n-1})$を求めるのは事実上不可能である。そのため、各単語の出力確率は$N-1$前の単語にのみ依存するという$N-1$重マルコフ性を仮定することでモデル構造を簡略化する。そして、単語の出力確率を単語列に渡って累積することにより、言語確率を計算する。このようなモデルをN-gram言語モデル[4][5]と呼ぶ。Nの値は3～4を用いることが多く、$N=3$の場合は特にTri-gramと呼ぶ*1。

N-gram言語モデルは大量のテキストデータに出現するN単語連鎖の頻度を用いて最尤推定することにより学習される。しかし、学習の際には低頻度のN単語連鎖（スパースネス）問題について考慮する必要がある。これに対処するための手法として、Back-off smoothing [30][31]という手法が広く用いられている。

2.3　WFSTに基づくモデル一体化

WFST [6][7]は有限状態オートマトンの一種であり、有限状態オートマトンは、あるアルゴリズムやモデルを状態遷移機械として表現して制御するために用いられる。有限状態オートマトンの最も単純なモデルは図3（a）に示すような、ある入力記号列を受理するか否かを判定する有限状態受理機械（FSA: Finite State Acceptor）である。例示したFSAは、{青い家}や{赤い車}という入力単語列及び{赤い家　青い車}という繰り返しを含むような入力単語列を受理する。図中、空語（empty word）ε *2を伴う状態遷移はε遷移と呼ばれ、入力無しでの状態遷移を示す。また、FSAの拡張モデルとして、図3（b）に示す有限状態トランスデューサー（FST: Finite State Transducer）があり、FSAの様に記号列を受理するだけでなく、入力記号列を別の記号列に変換する機能を有する。図3（b）は発音辞書をFSTとして表現したモデルあり、{a k a}とい

*1　$N=1$の場合はUni-gram、$N=2$の場合はBi-gramと呼ぶ。
*2　長さ0の特殊な記号であり、空文字列もしくはヌル文字列とも言う。空集合ϕを意味するものではない。



(a) Finite State Acceptor (FSA)

(b) Finite State Transducer (FST)

図3　FSAとFSTの例

う入力音素列を受理して {赤} という単語を出力する。また、FST の状態遷移の際に重み付けを行うことで、記号変換の起こりやすさ、起こりづらさを制御することが可能となる。このモデルを WFST と呼ぶ。

統計的音声認識における WFST は音響モデル、発音辞書、言語モデルを統合して、グラフ構造を持つ 1 つの巨大なネットワークとして表現されており、入力された特徴量ベクトル系列$\boldsymbol{O}$を直接最尤の単語列$\widehat{\boldsymbol{W}}$に変換するモデルとなっている。また、WFST は 3 つのモデルを単純に合成するのではなく、構造最適化を行うことにより、不要な経路や重複する経路を削除して軽量化を行っている。このような合成ネットワークモデルを用いることにより、デコーダの構造や探索アルゴリズムが簡略化されるという利点がある。WFST の合成、最適化アルゴリズム [6][32] や、デコーダにおける探索 [7][33] の詳細については文献を参照されたい。

3　深層学習に基づく音声認識

次に、深層学習 [9]–[11] に基づく音声認識について述べる。深層学習は機械学習技術の一種であり、生物の神経回路網を模した数理モデルであるニューラルネットワークを学習する方法及びその周辺技術の総称である。深層学習以前のニューラルネットワーク研究は 1957 年頃及び 1986 年頃に活発化したが、

1. 十分な演算能力を持つ計算機が確保できない
2. 効率的な学習アルゴリズムが確立されていない
3. 大量の学習データが利用できない

という理由により、大規模なニューラルネットワークを学習することができず、十分な性能を示すことができなかった。その後 2006 年頃に深層学習が登場し、

1. 汎用画像処理用演算プロセッサ（GPGPU: General Purpose Graphical Processing Unit）を用いた超高速並列演算と演算ライブラリの整備
2. 事前学習（Pre-training）や、確率的勾配降下法（SGD: Stochastic Gradient Descent）等の効率的な学習アルゴリズムの確立
3. 大量の学習データの利用

という技術革新により、多層構造をもつ大規模なニューラルネットワーク（DNN: Deep Neural Network）を学習することが可能となった。深層学習は、画像認識、音声認識等の様々な分野において従来技術をはるかにしのぐ性能を示し、瞬く間に普及した。

3.1　ハイブリッド型音声認識

深層学習の音声認識への導入は、統計的音声認識の一部を深層学習により得られたモデルに置き換えることから始まった。**2** にて述べたとおり、音声認識の背景には音響信号処理、機械学習等に基づく様々な技術があり、ある日突然それら全ての技術が深層学習に換装された訳ではない。1 つひとつの構成モジュールが見直されて、継続的な研究開発の過程で少しずつ深層学習によるモジュールに換装された。このように統計的音声認識と深層学習の混合による音声認識をハイブリッド型音声認識と呼ぶ。ハイブリッド型音声認識で主に換装されたモジュールは音響モデルと言語モデルであり、以降それぞれの概要について述べる。

3.1.1　深層学習に基づく音響モデル

従来の GMM-HMM 音響モデルは **2.1** にて述べたように、特徴量ベクトル系列$\boldsymbol{O}$から最尤の音素列$\widehat{\boldsymbol{S}}$を得る。より正確には、ある時刻tの特徴量ベクトル$\boldsymbol{o}_t$が、どの音素sのどの HMM 状態$q_{s,i}$に属するかを Viterbi アルゴリズムにより求める。深層学習による音響モデルでは、特徴量ベクトル$\boldsymbol{o}_t$を入力した際に、HMM 状態$q_{s,i}$に属する確率を要素にもつベクトル$\boldsymbol{q}_t$を直接出力するような DNN を学習して利用する [15]。ここで、ハイブリッド型音声認識の DNN 音響モデルでは入力特徴量ベクトル系列$\boldsymbol{O}=\{\boldsymbol{o}_0,\cdots,\boldsymbol{o}_t,\cdots,\boldsymbol{o}_T\}$と HMM 状態確率ベクトル系列$\boldsymbol{Q}=\{\boldsymbol{q}_0,\cdots,\boldsymbol{q}_t,\cdots,\boldsymbol{q}_T\}$の系列長が一致している必要がある。なお、このような枠組みは 1990 年代にすでに検討されていたが [34][35]、前述のとおり当時は高精度なニューラルネットワークを

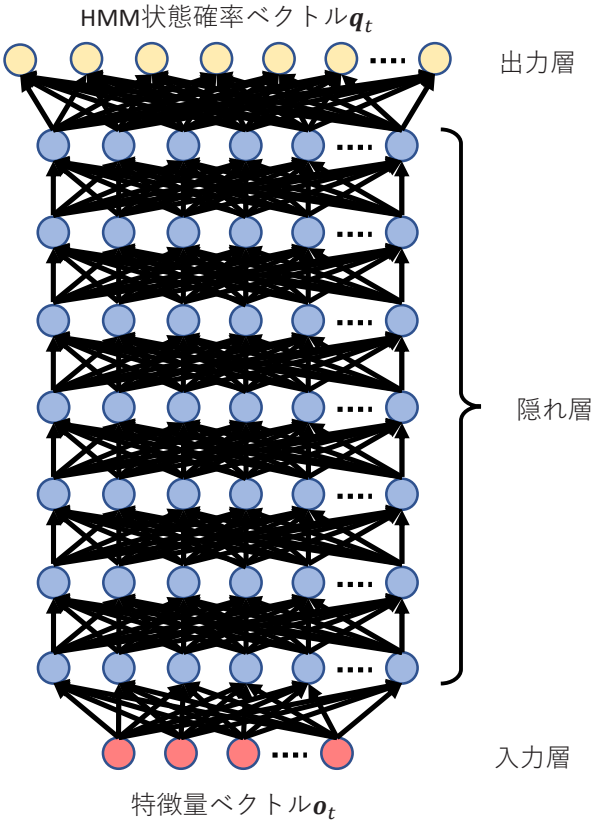
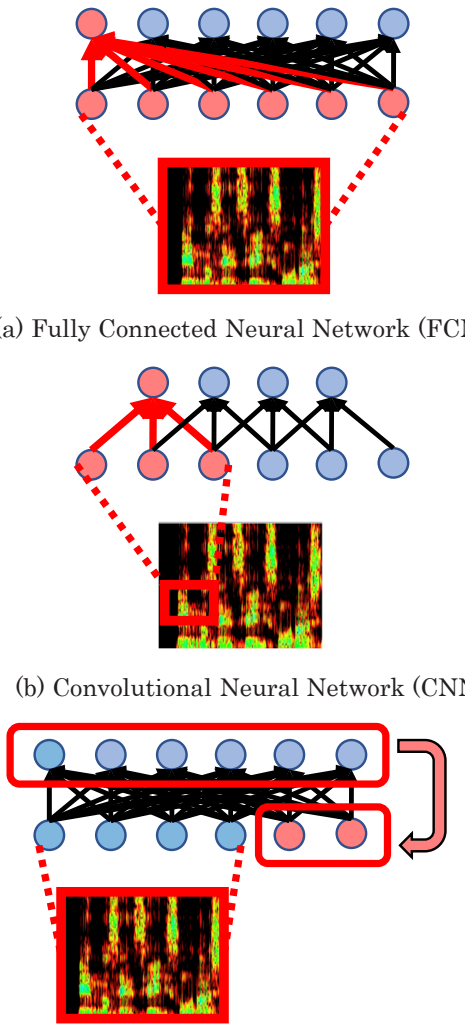


図 4　DNN 音響モデルの例

学習することができなかった。このため、GMM-HMM 音響モデルに取って代わることはなかったが、深層学習を用いた DNN 音響モデルの登場によりこれが覆された。

図 4 に DNN 音響モデルの一例を示す。図 4 の DNN は 7 層の隠れ層を有しており、特徴量ベクトル$\boldsymbol{o}_t$を入力すると隠れ層を順に伝搬していき、最終的に HMM 状態確率ベクトル$\boldsymbol{q}_t$を出力する。このような音響モデルを GMM-HMM 音響モデルと対比して、DNN-HMM 音響モデルと呼ぶことがある。深層学習以前の隠れ層数は多くとも 3 層程度が限度であったが、DNN では多数の隠れ層を有することが可能となった。また、各隠れ層は複数のノードから構成されている。深層学習以前は 100 程度が限度であったが、DNN では 1,000 ～ 2,000 のノードを有している。このような多くの隠れ層、ノードを有する DNN-HMM 音響モデルを用いることで、音声認識の性能が飛躍的に改善された。

図 4 に示した DNN-HMM 音響モデルは単純な構造となっているが、他の様々な構造を有するネットワークを用いることで音声認識性能を更に改善することができる。図 4 の DNN-HMM 音響モデルは、図 5 (a) の Fully Connected Neural Network (FCNN) を積み重ねることで構成されている。FCNN は層内のノードと前後の層のノードが全て結合された全結合ネットワークとなっており、入力された特徴量が全てのノードに伝搬する。一方、図 5 (b) の Convolutional Neural Network (CNN) [36]–[38] は小規模なカーネルを入力特徴量に畳み込んで情報伝達を行っており、ノード間を部分的に結合する部分結合ネットワークを構成する。すなわち、CNN は入力された特徴量から局所的な情報を抽出して伝搬することができる。FCNN では入力特徴量に含まれる全ての雑音の情報が伝播（でんぱ）されるのに対し、CNN では局所的な雑音の情報のみが伝搬されるため、その後のネットワークにおける雑音の影響を軽減することができる。そのため、CNN は雑音に対して頑健であるとされている。また、 図 5 (c) の Recurrent Neural Network (RNN) [39]–[41] は、隠れ層の出力を入力にフィードバックすることで、過去の情報を考慮することができ、音声信号のような時系列信号の解析、モデル化に有用である。RNN の発展形として、Long-Short Term Memory (LSTM) [42] がある。LSTM は記憶セル(Memory cell)とゲート機構 (Gating mechanism) という機能を有している。記憶セルには過去の情報が格納されており、その状態をゲート機構により制御する。ゲートには入力、忘却、出力の 3 種類があり、それぞれ記憶セルの更新、忘却、活用に相当する。また、過去から未来への順方向のネットワークだけでなく、未来から過去への逆方向のネットワークも取り込んだ Bi-directional RNN (LSTM) [43] も存在する。さらには CNN と LSTM を統合した Convolutional LSTM [44] や、CNN、LSTM、FCNN を積み重ねた複合的なネットワーク [45] も提案されており、DNN-HMM 音響モデルの構造は極めて多岐に渡る。



(a) Fully Connected Neural Network (FCNN)

(b) Convolutional Neural Network (CNN)

(c) Recurrent Neural Network (RNN)

図 5　様々なニューラルネットワーク構造

DNN-HMM 音響モデルの学習には、学習用のラベルデータとして入力特徴量ベクトル系列$\boldsymbol{O} = \{\boldsymbol{o}_0, \cdots, \boldsymbol{o}_t, \cdots, \boldsymbol{o}_T\}$に 1 対 1 で対応する HMM 状態系列が必要であり、この HMM 状態系列は別途学習した GMM-HMM 音響モデルを用いて、強制アライメント (Forced alignment) [46] という手法により得ることができる。学習時の損失関数には主に Cross entropy が用いられ、誤差逆伝搬法 (Back propagation) を用いて損失関数を最小化するように DNN の各パラメータを最適化する [9]–[11]。より発展的には State-level Minimum Bayes Risk (sMBR) [47]、Lattice Free-Maximum Mutual Information (LF-MMI) [48] 等の様々な基準での方法が確立されているが、紙面の関係から詳細については文献を参照されたい。

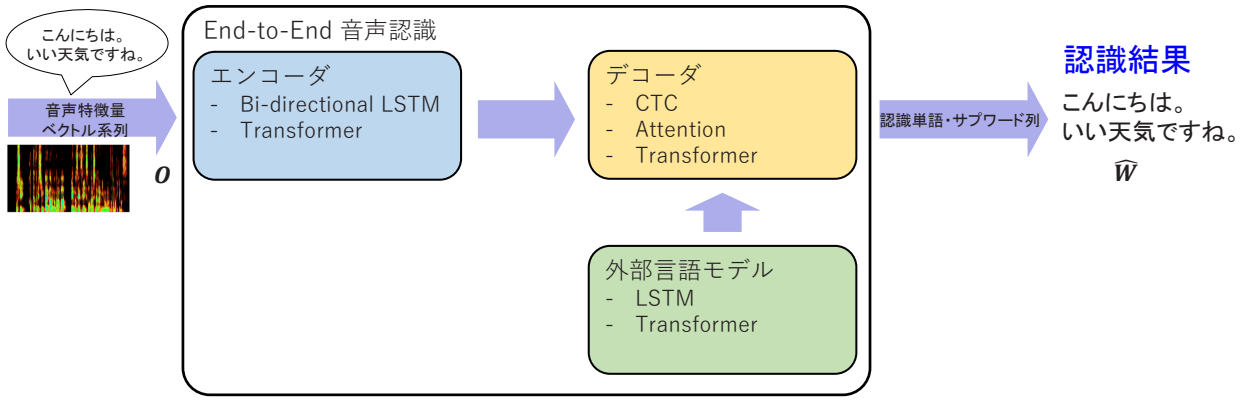


図 6　エンコーダ・デコーダネットワークによる E2E 音声認識

3.1.2　深層学習に基づく言語モデル

言語モデルにおいても深層学習が用いられる。RNN を用いて言語モデルを学習することにより、N-gram 言語モデルよりも長いコンテキストを考慮した処理が可能となる [40][41]。しかし、RNN 言語モデルを WFST の枠組みで直接導入することは実装上困難であるため、WFST の出力として得られた認識結果をリスコアリングすることにより音声認識の性能を改善することができる。

自然言語処理の研究分野では、極めて膨大なパラメータ数を持つ Bi-directional Encoder Representations from Transformers（BERT）[49] や、Generative Pre-trained Transformer-2（GPT-2）[50] という巨大ニューラル言語モデルの研究が盛んに行われている。RNN 言語モデルと同様に、BERT や GPT-2 も直接音声認識の枠組みで利用することは困難であるが、BERT による音声認識誤りの訂正、GPT-2 を特定の音声認識ドメイン（タスク）にファインチューニングした後に大量の文生成をして N-gram 言語モデルの学習に利用する等、間接的な利用の検討が進められている。特に後者はデータ拡張（Data augmentation）[51][52] と呼ばれており、学習データを何らかの方法（データ生成、雑音付加、一部欠損等）を用いて大幅に拡張し、モデル学習に利用する方法である。深層学習では大量の学習データが必要となるため、このような技術も同時に発展している。

3.2　End-to-End 音声認識

ハイブリッド型音声認識は、統計的音声認識の一部機能を深層学習モデルに置き換えることで実現されている。これに対して、式（1）の音声認識モデル $p(\boldsymbol{W}|\boldsymbol{O})$ を 1 つのニューラルネットワークで記述し、音声認識の問題を深層学習のみで解決しようという試みが注目を集めている。この試みは End-to-End（E2E）音声認識と呼ばれる [16]–[20][53][54]。統計的音声認識及びハイブリッド音声認識では性能改善を得るために、各構成モジュールを個別に改善し、個別に最適化することが行われてきた。これに対して E2E 音声認識は構成モジュールの改善は個別に行うものの、システムとしては 1 つのネットワークで記述されているため、システム全体の最適化を容易に行うことができる。またシステム構成としてもシンプルになる。

E2E 音声認識における最大の問題は、入力特徴量ベクトル系列 $\boldsymbol{O}=\{\boldsymbol{o}_0,\cdots,\boldsymbol{o}_t,\cdots,\boldsymbol{o}_T\}$ と，出力単語列 $\boldsymbol{W}=\{w_0,\cdots,w_n,\cdots,w_N\}$ の系列長が一致しないことである（基本的に入力に比べて出力の系列長が短い）。この問題に対処するため、入出力の系列長を調整するようなネットワーク構造が必要となる。E2E 音声認識では多くの場合、図 6 に示すようなエンコーダ・デコーダ（符号器・復号器）ネットワーク [55] にて構成される。図 6 においてエンコーダは、入力された特徴量ベクトル系列 $\boldsymbol{O}$ を音声認識のための適切な中間表現系列 $\boldsymbol{H}=\{\boldsymbol{h}_0,\cdots,\boldsymbol{h}_t,\cdots,\boldsymbol{h}_T\}$ に変換する役割を持ち、Bi-directional LSTM や Transformer [56] 等のネットワーク構造を有することが多い。デコーダはエンコーダ出力である中間表現系列 $\boldsymbol{H}$ を利用して、入出力の系列長を調整して出力する役割を持ち、Connectionist Temporal Classification（CTC）[57] 及び注意機構（Attention mechanism）[58] 等の手法が主に用いられる。

E2E 音声認識では音響モデルと言語モデルの区別が無く、利用できる学習データは基本的に音声データとその書き起こしテキストのみとなる。しかし、このような学習データは限られた量しか存在しないため、ある特定のドメインに適合することが困難である。このため統計的・ハイブリッド型音声認識で用いられるような従来の言語モデルを別途用意し、外部の知識を取り入れる方法が用いられている。外部言語モデルを導入する方法として Shallow fusion [59] があり、デコーダから得られる出力単語列の分布と、言語モデルの出力単語列の分布を重み付け平均することで、最終的な

出力単語列を得る。

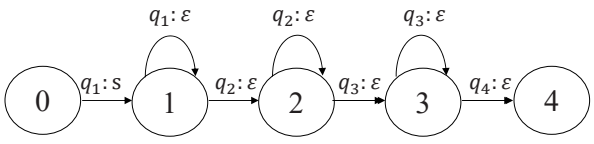
以下、E2E 音声認識の代表的なデコーダである CTC と Attention の概要について述べる。また、エンコーダ・デコーダネットワークではない、新たな方法である RNN-Transducer [60] についても簡潔に述べる。

3.2.1　CTC デコーダ

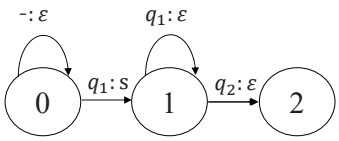
CTC デコーダ [61][62] では通常の出力シンボルに加えて「シンボルラベル無し」を意味するブランクシンボル ⊦⊣ を導入し、入力の一部をブランクシンボルに対応させることで、入力系列に比較して短い系列長の出力系列を出力する。系列長$T = 10$の入力系列が与えられた場合、出力系列 ⊦s e e d⊣ を得るための系列長$T = 10$の出力系列パターンは、

- ⊦s s e e e − e d − −⊣
- ⊦s e − − − e e − d −⊣
- ⊦− − s e e − e e d d⊣

等が考えられる。ここで ⊦s e e d⊣ は ⊦e⊣ が連続するので、それぞれを区別するためにブランクシンボル ⊦−⊣ を利用する。上記のいずれかのパターンから連続するシンボルをマージし、ブランクシンボル ⊦−⊣ を削除することで出力系列 ⊦s e e d⊣ を得ることができる。この様子を 3 状態 Left-to-Right HMM による GMM-HMM 音響モデルの FST 表現と比較すると、図 7 の様になる。図 7 (a) の Left-to-Right HMM では 3 つの HMM 状態q_1、q_2、q_3を遷移することでシンボル ⊦s⊣ を出力することが示されており、図 7 (b) の CTC ではブランクシンボル ⊦−⊣ が繰り返される中で、少なくとも 1 回状態q_1に遷移することでシンボル ⊦s⊣ を出力することが示されている。



(a) Left-to-Right HMM



(b) CTC

図 7　各モデルの FST 表現

DNN-HMM 音響モデルの学習では事前に強制アライメントで得た HMM 状態系列をラベルとして学習を行うが、CTC デコーダでは事前にこのようなアライメント情報を用意するのではなく、アライメント情報を探索しながら学習を行う必要がある。実際には Bi-directional LSTM エンコーダの出力系列$\boldsymbol{H}$と、HMM 学習に用いる Baum-Welch アルゴリズムに類似した Forward-Backward アルゴリズムを用いて学習を行う [19][62]。

3.2.2　Attention デコーダ

Attention デコーダ [63][64] は一般に順方向の LSTM で構成されており、エンコーダ出力系列$\boldsymbol{H}$と、過去に出力した単語列（接頭辞）$\{w_0, \cdots, w_{n-1}\}$を用いて次の出力単語w_nを予測する。図 8 に Attention デコーダの概要を示す。

図 8 において、Attention デコーダはエンコーダ出力系列$\boldsymbol{H}$をそのまま用いるのではなく、注意機構より得られた情報$\boldsymbol{c}_n$に変換しており、この情報は文脈ベク

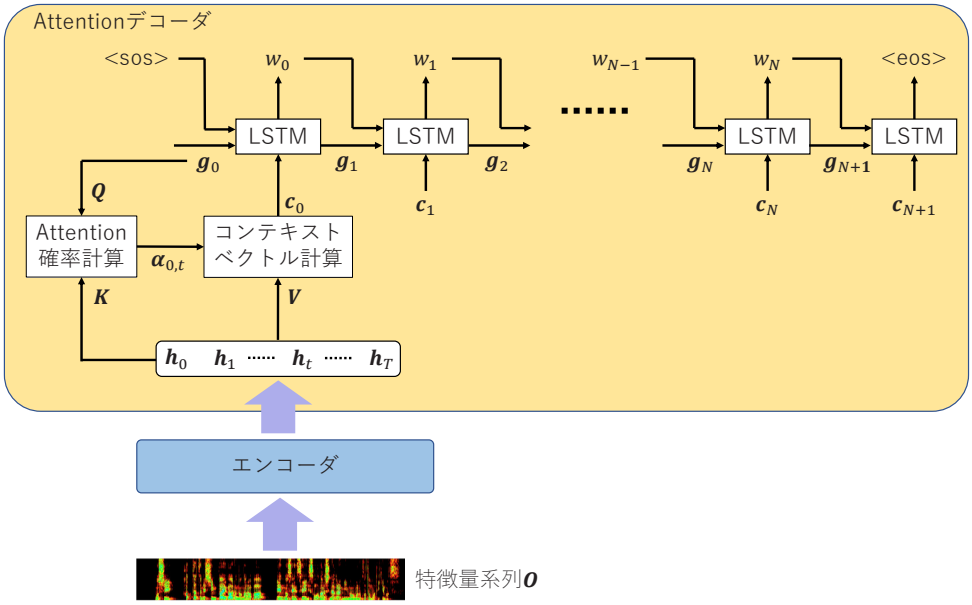


図 8　Attention デコーダの概要。<sos>、<eos> はそれぞれ文頭、文末のシンボルを示す。

トル（Context vector）と呼ばれる。注意機構は、n番目の単語w_nを予測するにあたり、エンコーダ出力系列$\boldsymbol{H}$の内、どの時刻の情報$\boldsymbol{h}_t$に注目すべきかをAttention確率$\alpha_{n,t}$に基づいて決定する手法であり、Query（$\boldsymbol{Q}$）、Key（$\boldsymbol{K}$）、Value（$\boldsymbol{V}$）と定義される3つの入力を受けて、文脈ベクトル$\boldsymbol{c}_n$を出力する（$\boldsymbol{K}$と$\boldsymbol{V}$は同じ情報源から得られ、これらの対をMemoryと呼ぶ）。文脈ベクトル$\boldsymbol{c}_n$は、

$$\boldsymbol{c}_n = \sum_t \alpha_{n,t} \cdot \boldsymbol{h}_t \tag{4}$$

により与えられ、上式の$\boldsymbol{h}_t$が$\boldsymbol{V}$に相当し、$\alpha_{n,t}$をAttention確率と呼ぶ。Attention確率$\alpha_{n,t}$の計算方法には幾つかの方法があるが、最も一般的なのは、デコーダLSTMの入力$\boldsymbol{g}_n$（$\boldsymbol{Q}$に相当）と、エンコーダ出力$\boldsymbol{h}_t$（$\boldsymbol{K}$に相当）との内積を求める手法であり、次式により与えられる。

$$\alpha_{n,t} = \frac{\exp(\boldsymbol{g}_n^{\mathrm{T}} \boldsymbol{h}_t)}{\sum_{t'} \exp(\boldsymbol{g}_n^{\mathrm{T}} \boldsymbol{h}_{t'})} \tag{5}$$

Attentionデコーダは CTC デコーダと異なり学習時に特別な方法を必要としない。また、CTC デコーダとAttention デコーダを統合したハイブリッド方式 [65] も提案されており、それぞれのデコーダを単体で利用するよりも高い音声認識性能が得られることが報告されている。

注意機構を用いたモデルの発展形として、自然言語処理や機械翻訳の分野で利用されているTransformerがある [56]。Transformer はエンコーダ、デコーダともに注意機構が用いられており、特に入力を幾つかのブロックに分割する Multi-head attention と、$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$を同一の情報源から得る Self-attention を有することが特徴的である。また、Transformer の一部を CNN に置き換えた Conformer [66] も提案されており、Transformer に比べて性能改善が得られることが報告されている。

3.2.3　RNN-Transducer モデル

RNN-Transducer [60] は、CTC デコーダと同様に、ブランクシンボルを用いて拡張された出力シンボルの推定を、接頭辞の情報を用いて実施するモデルとなっている。また、入力と出力が同期する設計となっており、処理遅延の少ないストリーミング音声認識向けの手法である。

図9はRNN-Transducerの概要図を示しており、特徴量ベクトル系列$\boldsymbol{O}$から中間表現$\boldsymbol{h}_t$を出力するエンコーダ、接頭辞$\{w_0, \cdots, w_{n-1}\}$から対応する言語情報ベクトル$\boldsymbol{p}_n$を予測して出力するPrediction network、それぞれの結果を統合するJoint networkで構成される。Joint networkは、入力としてT通りの値を持つ中間表現ベクトル$\boldsymbol{h}_t$と、N通りの値を持つ言語情報ベクトル$\boldsymbol{p}_n$を$T \times N$通りの組み合わせで受け取り、tとnそれぞれの場合において次に出力すべき単語の確率分布$p(\boldsymbol{W}|t,n)$を出力する。

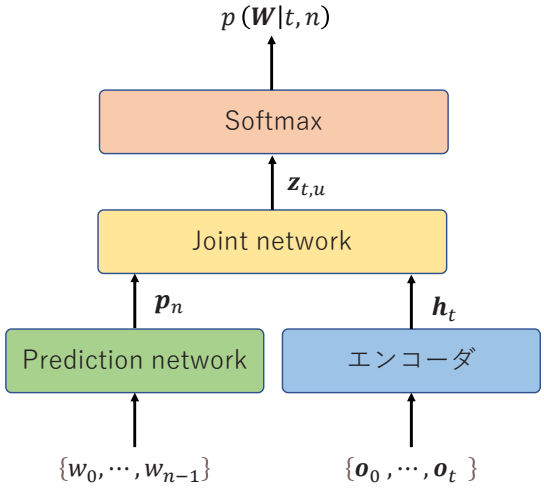


図9　RNN-Transducerの概要

RNN-Transducerでは一般にエンコーダ、Prediction networkともに順方向のLSTMにより構成され、Joint networkはFCNNにより構成される。最新の研究ではTransformer、Conformer等の注意機構を用いた手法も提案されている [67]。なお、RNN-Transducerの学習はCTCデコーダと同様の方法にて行われる。

4　NICTにおける取組

上記に述べた音声認識の技術発展において、NICTにおいても第一線で研究開発を推進し、多くの研究成果を挙げて技術発展に貢献してきた。以下、NICTにおける音声認識の研究開発及びその成果の展開について述べる。

4.1　多言語音声翻訳技術の展開

NICTでは多言語音声翻訳技術の社会実装を至上命題として研究開発を推進している。その1つの成果としてモバイル端末における多言語音声翻訳アプリVoiceTra® [21] があり、音声翻訳技術の社会実装を行うための実証実験の名目で無償公開している。音声翻訳の要素技術は音声認識、機械翻訳、音声合成であり、ユーザーの音声を受け取る音声認識は音声翻訳の入り口にあたる。そのため音声認識は極めて重要な役割を担っており、特に注力して研究開発を推進している。

VoiceTra® とその音声翻訳エンジンは、総務省主導で遂行されたグローバルコミュニケーション計画（GCP）[68] の目標である「音声翻訳技術の社会実装」において、中核的な役割を果たした。現在は、「多言語同

時通訳技術の社会実装」を目標としたGCP 2025 [69]が遂行されており、GCP2025においてもNICTの技術がその中核を担う予定である。なお、GCP、GCP2025及びNICTの技術展開の詳細については本特集号 **2–4** の記事を参照されたい。

GCP2025は講演、会議を対象とした同時通訳を目標としており、その延長として音声認識を用いた自動講演録、会議録生成システムの開発を推進している。本システムでは音声認識のみを用いるのではなく、本特集号 **2–2–5** で解説する言語識別・話者識別技術と組み合わせることにより詳細な講演録、会議録を自動生成することを目指している。

4.2　多言語音声認識技術の研究開発

VoiceTra® では様々な言語による音声入力を受けつけるため、必然的に多言語の音声認識が必要となる。そのため、NICTではアジア言語を中心として以下の19言語の音声認識を開発し、実装している（2022年8月現在）。今後、イタリア語、ドイツ語、ヒンディー語に対応する予定である。

- 主要4言語：日本語、英語、中国語（簡体字）、韓国語
- アジア言語：台湾華語（繁体字）、インドネシア語、タイ語、ベトナム語、ミャンマー語、フィリピン語、クメール語、ネパール語、モンゴル語
- ヨーロッパ言語：スペイン語、フランス語、ロシア語、ウクライナ語
- その他：アラビア語、ブラジルポルトガル語

各言語の音声認識において、主要4言語については人間レベルの音声認識性能を達成しており、他の言語においても実用、準実用レベルの音声認識を達成している*3。英語に関しては、音声翻訳に関する国際ワークショップIWSLTの音声認識チャレンジ（TED講演の英語音声認識タスク）において、2012年から2014年までの3年間連続で性能1位を獲得した[70]–[72]。

現状、NICTにおける音声認識の研究開発は、ハイブリッド型音声認識とE2E音声認識の2ラインで実施している。前者のハイブリッド型音声認識は、VoiceTra® を含む実システムのための研究開発であり、日々着実に技術改善している。NICTでは多種多様な音声認識タスクに対応するため、新単語登録等の様々な機能を実現する周辺ツールも合わせて開発している。現在のE2E音声認識の枠組みでは新単語登録等の機能実装が容易ではなく、実システムの開発においては、ハイブリッド型音声認識の需要はいまだ高いと考え、継続的に研究開発を行っている。一方、E2E音声認識は基礎研究レベルでの検討段階にあり、一部言語ではプロトタイプシステムを開発して動作を確認している。今後、多言語化及び実システムの開発に加えて、ハイブリッド型音声認識と同等の周辺機能実装を行い、段階的にハイブリッド型音声認識から移行することを検討している。

4.3　耐雑音性・耐残響性の確立

VoiceTra® はモバイルデバイス上のアプリであるため、屋内外での利用が想定される。特に屋外では音声入力時に周囲の雑音が混入する場合があり、さらに口（音源）とマイクの距離が遠い場合、雑音だけでなく反射音によって生じる残響も混入する可能性が高い。このような環境下では、雑音や残響の影響により音声認識性能が著しく劣化する。この問題については、データ拡張を行うことにより対処している。すなわち、種々の雑音や残響を人工的に付加した学習用音声データを大量生成し、音響モデル学習を行っている。また、音声データに対して意図的にランダムなデータ欠損を生じさせる手法[51]も導入して、音声認識の耐雑音性・耐残響性を確立している。

基礎研究レベルでは、多チャネル音声入力を考慮した手法と、複数の単一チャネル雑音除去手法の出力をマルチストリーム入力として受け付ける手法について提案を行った。前者は、Beam-formerと呼ばれる多チャネル信号処理技術をニューラルネットワークで実現し、音声認識のDNN音響モデルと統合して全体最適化を行う手法である[73][74]。この手法は、Beam-formerとDNN音響モデルを1つのニューラルネットワークで記述する部分的なE2E手法となっている。後者についてはStream-wise transformer [75]という手法を提案した。この手法は、複数の雑音除去手法からの出力で構成されたマルチストリーム入力に対して注意機構を適用し、雑音環境ごとに最適なストリーム（雑音除去処理）を自動選択可能とする手法である。今後、これら手法の実システムへの実装について検討を行う。

4.4　自動字幕表示システムの研究開発

NICTでは、総務省の「平成30年度情報通信利用促

*3　音声認識の性能基準は以下の定義に従っている。
- 人間レベル：音声認識結果を読んで問題なく理解できる
- 実用レベル：軽微な誤りがあるが音声認識結果を読んで十分に理解できる
- 準実用レベル：誤りがあるが音声認識結果を読んである程度理解できる
- 実験レベル：誤りが多く音声認識結果を読んで理解するのが難しい
- 試作レベル：誤りが多く音声認識結果を読んで理解するのが極めて困難

進支援事業費補助事業（聴覚障害者放送視聴支援緊急対策事業）」[76] に採択されたことを受け、放送番組を始めとする様々な映像メディアに対する自動字幕表示システムの研究開発を推進している。本事業における実証実験では、複数の放送事業者が実際に放送した番組上で自動字幕表示システムを稼働させてリアルタイムでの字幕表示実験を行った。実証実験後に実施したアンケート調査の結果、ニュース番組等では十分実用に足る性能であるとのコメントが数多くあり、高い評価を得ることができた。現在、幾つかの機関との共同研究を通じて実用化に向けた検討を行っている。

5　あとがき

本稿では、音声認識の技術動向について俯瞰し、統計的音声認識、ハイブリッド型音声認識、E2E 音声認識への移り変わりと、それらの代表的な技術について解説を行った。また、このような世界的な技術発展における NICT の取組について紹介した。今後、音声認識技術はますます発展していくことが予想されるが、NICT においてもそれに追随して、今後も第一線での研究開発を推進する予定である。

【参考文献】

1 L. Rabiner and C. Schmidt, “Application of dynamic time warping to connected digit recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.28, issue 4, pp.377–388, Aug. 1980.
2 C. M. Bishop, “Pattern recognition and machine learning,” Springer, 2006.
3 X. D. Huang, Y. Ariki, and M. A. Jack, “Hidden Markov models for speech recognition,” Edinburgh University Press, 1990.
4 H. Ney, U. Essen, and R. Kneser, “On structuring probabilistic dependences in stochastic language modelling,” Computer Speech & Language, vol.8, issue 1, pp.1–38, Jan. 1994.
5 R. Rosenfeld, “The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation,” Proceedings of ARPA Spoken Language Systems Technology Workshop, pp.47–50, Jan. 1995.
6 堀 貴明，塚田 元，“重み付き有限状態トランスデューサによる音声認識，”情報処理学会誌，45 巻，10 号，pp.1020–1026，Oct. 2004.
7 T. Hori, C. Hori, Y. Minami, and A. Nakamura, “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol.15, issue 4, pp.1352–1365, May 2007.
8 F. Jelinek, “Statistical methods for speech recognition (Language, speech, and communication),” MIT Press, 1998.
9 Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, vol.521, no.7553, pp.436–444, May 2015.
10 I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” MIT Press, 2016.
11 麻生 英樹，安田 宗樹，前田 新一，岡野原 大輔，岡谷 貴之，久保 陽太郎，ボレガラ ダヌシカ，神嶌 敏弘，“深層学習 ― Deep learning，”近代科学社，2015.
12 D. Yu and L. Deng, “Automatic speech recognition: A deep learning approach,” Springer, 2015.
13 S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, “New era for robust speech recognition - Exploiting deep learning,” Springer, 2017.
14 U. Kamath, J. Liu, and J. Whitaker, “Deep learning for NLP and speech recognition,” Springer, 2019.
15 久保 陽太郎，“音声認識のための深層学習，”人工知能，29 巻，1 号，pp.62–71，Jan. 2014.
16 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，73 巻，1 号，pp.31–38，Jan. 2017.
17 渡部 晋治，堀 貴明，“音声言語理解のための音声認識，”電子情報通信学会誌，vol.101，no.9，pp.885–890，Sept. 2018.
18 高島 遼一，“Python で学ぶ音声認識，”インプレス，2021.
19 久保 陽太郎，“機械学習による音声認識，”コロナ社，2021.
20 渡部 晋治，久保 陽太郎，“深層学習が支える音声認識技術，”電子情報通信学会誌，vol.105，no.5，pp.392–396，May 2022.
21 多言語音声翻訳アプリ VoiceTra, https://voicetra.nict.go.jp
22 河原 達也，“音声認識システム，”オーム社，2006.
23 L. E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov chains,” The Annals of Mathematical Statistics, vol.37, no.6, pp.1554–1563, Dec. 1996.
24 村上 仁一，“Baum-Welch アルゴリズムの動作と応用例，”IEICE Fundamentals Review, vol.4, no.1, pp.48–56, Jan. 2010.
25 A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” IEEE Transactions on Information Theory, vol.13, issue 2, pp.260–269, April 1967.
26 G. D. Forney, “The Viterbi algorithm,” in Proceedings of the IEEE, vol.61, issue 3, pp.268–278, March 1973.
27 S. J. Young and P. Woodland, “The use of state tying in continuous speech recognition,” Proceedings of Eurospeech ’93, pp.2203–2206, Sept. 1993.
28 J. R. Bellegarda and D. Nahamoo, “Tied mixture continuous parameter models for large vocabulary isolated speech recognition,” Proceedings of ICASSP ’89, pp.13–16, May 1989.
29 鷹見 淳一，嵯峨山 茂樹，“逐次状態分割による隠れマルコフ網の自動生成，”電子情報通信学会論文誌，J76-DII, vol.10, pp.2155–2164, Oct. 1993.
30 S. M. Katz, “Estimation of probabilities from sparse data for the language model component of a speech recognizer,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.35, issue 3, pp.400–401, March 1987.
31 川端 豪，“二項事後分布に基づく N-gram 記号連鎖確率の推定，”日本音響学会誌，61 巻，8 号，pp.441–447，Aug. 2005.
32 大西 翼，ポール ディクソン，岩野 公司，古井 貞熙，“WFST 音声認識デコーダにおける on-the-fly 合成の最適化処理，”電子情報通信学会論文誌，J92-DII, vol.7，pp.1026–1035，July 2009.
33 ポール ディクソン，堀 智織，柏岡 秀樹，“SprinTra WFST 音声デコーダ開発について，”情報通信研究機構研究報告，58 巻，3/4 号，pp.13–18，Sept./Dec. 2012.
34 H. A. Bourlard and N. Morgan, “Connectionist speech recognition: A hybrid approach,” Kluwer Academic Publishers, Oct. 1993.
35 S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, “Connectionist probability estimators in HMM speech recognition,” IEEE Transactions on Speech and Audio Processing, vol.2, issue 1, pp.161–174, Jan. 1994.
36 O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol.22, issue 10, pp.1535–1545, Oct. 2014.
37 T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR,” Neural Networks, vol.64, pp.39–48, April 2015.
38 Y. Zhang, W. Chan, and N. Jaitly, “Very deep convolutional networks for end-to-end speech recognition,” Proceedings of ICASSP ’17, pp.4845–4849, June 2017.
39 N. Kanda, M. Tachimori, X. Lu, and H. Kawai, “Training data pseudo-shuffling and direct decoding framework for recurrent neural network based acoustic modeling,” Proceedings of ASRU ’15, pp.13–17, Dec. 2015.
40 Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” The journal of machine learning research, vol.3, pp.1137–1155, March 2003.
41 T. Mikolov, L. Karafiat, L. Burget, J. Cernocky and S. Khudanpur, “Recurrent neural network-based language model,” Proceedings of Interspeech ’10, pp.1045–1048, Sept. 2010.
42 H. Sak, A. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” Pro-

ceedings of Interspeech ’14, pp.338–342, Sept. 2014.
43 A. Graves, N. Jaitly, and A. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM,” Proceedings of ASRU ’13, pp.273–278, Dec. 2013.
44 X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” Proceedings of NIPS ’15, pp.802–810, Dec. 2015.
45 T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” Proceedings of ICASSP ’15, pp.4580–4584, April 2015.
46 S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. Woodland, and C. Zhang, “The HTK book (version 3.5a),” https://www.danielpovey.com/files/htkbook.pdf, Dec. 2015.
47 K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” Proceedings of Interspeech ’12, pp.2345–2349, Aug. 2013.
48 D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free,” Proceedings of Interspeech ‘16, pp.2751–2755, Sept. 2016.
49 J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of NAACL-HLT ’19, pp.4171–4186, June 2019.
50 A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” Technical report of OpenAI, June 2018.
51 D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” Proceedings of Interspeech ’19, pp.2613–2617, Sept. 2019.
52 C. Du, H. Li, Y. Lu, L. Wang, and Y. Qian, “Data augmentation for end-to-end code-switching speech recognition,” Proceedings of SLT ’21, pp.194–200, Jan. 2021.
53 林 知樹,“End-to-End 音声処理の概要と ESPnet2 を用いたその実践,”日本音響学会誌，76 巻，12 号，pp.720–729，Dec. 2020.
54 河原 達也,“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 ―,”日本音響学会誌，74 巻，7 号，pp.381–386，July 2018.
55 I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” Proceedings of NIPS ’14, vol.27, Dec. 2014.
56 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proceedings of NeurIIPS ’17, Dec. 2017.
57 A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” Proceedings of ICML ’06, pp.369–376, June 2006.
58 D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Proceedings of ICLR ’15, May 2015.
59 S. Toshniwal, A. Kannan, C. Chiu, Y. Wu, T. N. Sainath and K. Livescu, “A comparison of techniques for language model integration in encoder-decoder speech recognition,” Proceedings of SLT ’18, pp.369–375, Dec. 2018.
60 Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, “Streaming end-to-end speech recognition for mobile devices,” in Proceedings of ICASSP ’19, pp.12-17, May 2019.
61 A. Graves and Nav. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” Proceedings of ICML ’2014, pp.1764–1772, June 2014.
62 H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” Proceedings of ICASSP ’15, pp.4280–4284, April 2015.
63 J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” Proceedings NIPS ’15, pp.577–585, Dec. 2015.
64 N. Moritz, T. Hori, and J. L. Roux, “Triggered attention for end-to-end speech recognition,” Proceedings of ICASSP ’19, pp.5666-5670, May 2019.
65 S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol.11, no.8, pp.1240–1253, Dec. 2017.
66 A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and, R. Pang, “Convolution-augmented transformer for speech recognition,” Proceedings of Interspeech ’20, pp.5036–5040, Oct. 2020.
67 Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, “Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss,” Proceedings of ICASSP ’20, pp.7829–7833, May 2020.
68 総務省,“グローバルコミュニケーション計画,”https://www.soumu.go.jp/main_content/000285578.pdf, April 2014.
69 総務省,“グローバルコミュニケーション計画 2025,”https://www.soumu.go.jp/main_content/000678485.pdf, March 2020.
70 H. Yamamoto, Y. Wu, C. Huang, X. Lu, P. R. Dixon, S. Matsuda, C. Hori, and H. Kashioka, “The NICT ASR system for IWSLT2012,” Proceedings of IWSLT ’12, Dec. 2012.
71 C. Huang, P. R. Dixon, S. Matsuda, Y. Wu, X. Lu, M. Saiko, and C. Hori, “The NICT ASR system for IWSLT 2013,” Proceedings of IWSLT ’13, Dec. 2013.
72 P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko and C. Hori, “The NICT ASR system for IWSLT 2014,” Proceedings of IWSLT ’14, Dec. 2014.
73 Masakiyo Fujimoto, “Factored deep convolutional neural networks for noise robust speech recognition,” Proceedings of Interspeech ’17, pp.3837–3841, Aug. 2017.
74 Masakiyo Fujimoto and Hisashi Kawai, “Comparative evaluations of various factored deep convolutional RNN architectures for noise robust speech recognition,” Proceedings of ICASSP ’18, pp.4829–4833, April 2018.
75 Masakiyo Fujimoto and Hisashi Kawai, “Noise robust acoustic modeling for single-channel speech recognition based on a stream-wise transformer architecture,” Proceedings of Interspeech ’21, pp.281–285, Sept. 2021.
76 総務省,“聴覚障害者放送視聴支援緊急対策事業,”https://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000228.html, March 2019.

藤本 雅清　（ふじもと まさきよ）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声音響信号処理、音声認識、機械学習
【受賞歴】
2003 年　日本音響学会 第20回栗屋潔学術奨励賞
2011 年　情報処理学会 2010年度（平成22年度）山下記念研究賞
2015 年　IEEE ASRU，’15 Best Paper Award Honorable Mention