日本音響学会 編

音響テクノロジーシリーズ 24

機械学習による音声認識

博士（工学） 久保　陽太郎　著

コ ロ ナ 社

発刊にあたって

音響テクノロジーシリーズは 1996 年に発刊され，以来 20 年余りの期間に 19 巻が上梓された。このような長期にわたる刊行実績は，本シリーズが音響学の普及に一定の貢献をし，また読者から評価されてきたことを物語っているといえよう。

この度，第 5 期の編集委員会が立ち上がった。7 名の委員とともに，読者に有益な書籍を刊行し続けていく所存である。ここで，本シリーズの特徴，果たすべき役割，そして将来像について改めて考えてみたい。

音響テクノロジーシリーズの特徴は，なんといってもテーマ設定が問題解決型であることであろう。東倉洋一初代編集委員長は本シリーズを「複数の分野に横断的に関わるメソッド的なシリーズ」と位置付けた。従来の書籍は学問分野や領域そのものをテーマとすることが多かったが，本シリーズでは問題を解決するために必要な知見が音響学の分野，領域をまたいで記述され，さらに多面的な考察が加えられている。これはほかの書籍とは一線を画するところであり，歴代の著者，編集委員長および編集委員の慧眼の賜物である。

本シリーズで取り上げられてきたテーマは時代の最先端技術が多いが，第 4 巻「音の評価のための心理学的測定法」のように汎用性の広い基盤技術に焦点を当てたものもある。本シリーズの役割を鑑みると，最先端技術の体系的な知見が得られるテーマとともに，音の研究や技術開発の基盤となる実験手法，測定手法，シミュレーション手法，評価手法などに関する実践的な技術が修得できるテーマも重要である。

加えて，古典的技術の伝承やアーカイブ化も本シリーズの役割の一つとなろう。例えば，アナログ信号を取り扱う技術は，技術者の高齢化により途絶の危

機にある。ディジタル信号処理技術がいかに進んでも，ヒトが知覚したり発したりする音波はアナログ信号であり，アナログ技術なくして音響システムは成り立たない。原理はもちろんのこと，ノウハウも含めて，広い意味での技術を体系的にまとめて次代へ継承する必要があるだろう。

コンピュータやネットワークの急速な発展により，研究開発のスピードが上がり，最新技術情報のサーキュレーションも格段に速くなった。このような状況において，スピードに劣る書籍に求められる役割はなんだろうか。それは上質な体系化だと考える。論文などで発表された知見を時間と分野を超えて体系化し，問題解決に繋がる「メソッド」として読者に届けることが本シリーズの存在意義であるということを再認識して編集に取り組みたい。

最後に本シリーズの将来像について少し触れたい。そもそも目に見えない音について書籍で伝えることには多大な困難が伴う。歴代の著者と編集委員会の苦労は計り知れない。昨今，書籍の電子化についての話題は尽きないが，本文の電子化はさておき，サンプル音，説明用動画，プログラム，あるいはデータベースなどに書籍の購入者がネット経由でアクセスできるような仕組みがあれば，読者の理解は飛躍的に向上するのではないだろうか。今後，検討すべき課題の一つである。

本シリーズが，音響学を志す学生，音響の実務についている技術者，研究者，さらには音響の教育に携わっている教員など，関連の方々にとって有益なものとなれば幸いである。本シリーズの発刊にあたり，企画と執筆に多大なご努力をいただいた編集委員，著者の方々，ならびに出版に際して種々のご尽力をいただいたコロナ社の諸氏に厚く感謝する。

2018 年 1 月

音響テクノロジーシリーズ編集委員会

編集委員長　飯田　一博

ま　え　が　き

音声認識は夢のテクノロジーである。人の言葉を聞き取り理解する技術は，人間の真のパートナーとなるべき機械もしくはロボットを実現する鍵となる技術である。

音声認識には，「人の音声をテキストに変換する」という大まかな定義があるが，入力となる音声や出力となるテキストの多様性から，問題のスコープを正確に特定することが難しい。初期の音声認識は，特定の話者の音声に限って，コマンド発話（「はい」か「いいえ」かなど）のみを認識できる程度の単純なものであった。また，そのような単純な機能であっても，その実現には当時の技術の粋を結集しなければならなかった。音声認識は，つねに人間と比べられるという性質から，その可用性について厳しい目で見られ続け，「未完成」であると過小評価されやすい技術であった。

翻って現在，スマートスピーカやスマートフォンのような製品とともに，音声認識を利用した情報家電が人々の暮らしの中に取り込まれつつある。これまでも音声認識を利用した製品やサービスは多く存在したが，最新の機械学習技術を用いて，より高い精度が達成されたことにより，音声認識はこれまで以上に身近になった。音声認識はついに，これまでの厳しい評価から抜け出しつつあるように見える。

このように急速に一般化しつつある音声認識であるが，人々の要求はいまも高度化し続けている。音声認識が身近になることで，「どのような状況で」「だれが」「どのようなことを」話しても認識できるようになることの重要性が，これまでより高まってきている。また，ほぼ人間と同精度での認識が可能になった現在，人間を超える認識精度への期待も高まりつつある。本書は，そのよう

な期待に応えうる未来の技術を切り拓くために，必要な知識を学ぶためのテキストである。

本書の主たる想定読者は，この分野に携わる技術者，研究者，およびこの分野の研究を始めようとする学生である。「この分野」の示すところは，なるべく広い範囲にわたるように留意した。例えば，音楽の情報処理は，音響信号の意味論を分析するという点で，多分に音声認識と重なり合う。また，音声認識は，音声を入力とする自然言語生成技術の 1 つであると考えることもできる。映像の理解に関して，音声トラックの認識に音声認識が直接利用される場合もあれば，音声認識の技術を拡張して，映像信号のテキスト化を行うといったような，より本質的な拡張もありうる。音声認識研究がさまざまな分野の基礎技術の上に成り立っているのと同様に，その果実である音声認識技術もさまざまな応用領域において利用可能であると信じている。

本書の執筆にあたって，さまざまな研究者から貴重な意見をいただいた。中村篤教授（名古屋市立大学），堀貴明博士（Mitsubishi Electric Research Laboratories）に謝意を示したい。著者が所属する Google 合同会社の同僚からも，さまざまな意見を頂戴した。大西翼博士，苅田成樹氏に謝意を示したい。加えて，本書を書くきっかけをくださった大川茂樹教授（千葉工業大学）に改めて謝意を表したい。

2021 年 3 月

久保　陽太郎

目　　次

1. 本書の目的と事前知識

2. 機械学習による予測

3. 有限状態トランスデューサ

4. 音声認識システム

5. 音響モデル

6. 言語モデル

7. 大語彙連続音声認識

8. 深層学習の発展

1 本書の目的と事前知識

本章では，本書の目的と，本書を読み進めるにあたって必要な事前知識を解説する。本章では，まず 1.1 節で本書の目的を述べる。つぎに，1.2 節では本書の構成を章ごとに説明する。1.3 節では本書で用いる数式の表記法について解説し，1.4 節では本書で用いる範囲で確率論の解説を行う。

1.1 本書の目的

本書では，音声認識システムを構成する技術について，**機械学習**（machine learning）の観点から解説する。人の声をテキストに変換する音声認識システムは非常に長い間研究されてきており，その背後にはさまざまな要素技術のエッセンスが詰まっている。多くの人がいとも簡単に行う音声の聞き取りは，コンピュータにとっては複雑であり，多分野にわたる理論と技術を，1 つのシステムに集約することで初めて可能になる処理である。

音声認識が現在のように大規模で複雑なシステムへと成長してきた背景には，問題そのものが難しくなってきたという歴史がある。一口に「音声認識」という言葉で表される問題には，数字の聞き分けのみができるというきわめて初歩的なレベルから，人間でも聞き取りが難しい，雑談の一言一句を聞き取るレベルまである。コンピュータ技術の一般化とともに，この分野への社会の期待と要求が高まる中，より難しい問題設定での音声認識を実現する技術が進展してきた。

図 1.1 に，音声認識の数ある技術的課題のうちのいくつかを示す。それぞれの課題にはそれに応じた達成目標があり，現代の技術であれば容易に達成でき

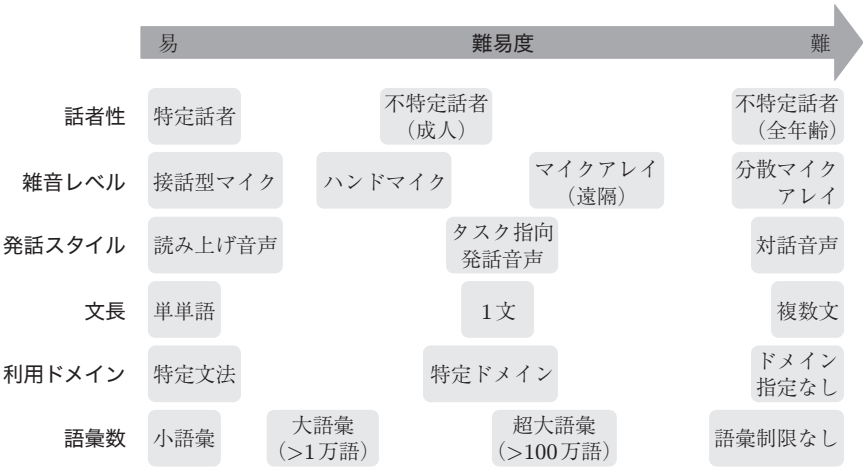

図 1.1　音声認識の達成目標と難易度

るものから，現代の技術でも困難なものまである。音声認識技術は，この図の一番左に並んだ難易度が「易」の目標を達成するところから始まり，それぞれの側面において，より難しい目標を達成するように進化してきた。そして，この進化の過程で，音声認識の理論とそれを実現するシステムは複雑化し続けてきた。今日の音声認識技術は，さまざまな要素技術の複合体である。

先述したように，本書は機械学習の言葉で音声認識を理論的側面から解説することを第一の目的とする。とはいえ，音声認識をシステムとして詳細に記述しようとする際に，機械学習の言葉のみでは十分ではない。そこで，本書では，システム記述の共通言語として**有限状態トランスデューサ**（finite-state transducer; FST）を用い，これら 2 つの視点から音声認識システムを俯瞰し，理論と実装の双方から体系的に解説することを試みる。

本書の 1 つ目のキーワードである「機械学習」は，データからそれを再現するアルゴリズムを得る技術の総称である。例えば，あるプログラムの入力と出力のペアを “(入力, 出力)” のように書くとして，“(0, false), (2, false), (3, true), (5, true)” というような入出力例があるとする。人間がこのデータを見ると，おそらくこのプログラムは入力が 3 以上なら true を返す関数かもしれないと予想する

ことができる。機械学習はそういった仮説を機械的に得る手法の総称である。有限の入出力例に対して，それらを完璧に再現するアルゴリズムは無限にある。この例では「入力が奇数のときに true を返す関数である」という別の仮説を立てることもできる。どの機械学習技術をどのように用いるかによって，どのような仮説が選ばれるかが決まる。したがって，さまざまな機械学習アルゴリズムの挙動について定性的な理解を持つことは，音声認識技術の研究では非常に重要である。

本書のもう 1 つのキーワードである「有限状態トランスデューサ」は，系列入力と系列出力を持つシステムの記述法の 1 つである。系列とは，リストや文字列のような長さの定まっていない複数要素を持つ構造のことである。音声認識器もまた，系列を入力として，受け取り単語の系列を出力する機械と見なすことができ，したがって有限状態トランスデューサで記述できる。現代の音声認識技術の多くの要素は有限状態トランスデューサで説明できるため，有限状態トランスデューサは音声認識のアルゴリズムを共通の枠組みで記述する上で重要である。

1.2 本 書 の 構 成

図 1.2 に本書の構成を示す。図中の矢印は前提知識の依存関係を示し，例えば，6 章には 2 章と 3 章で紹介した事柄が前提知識として用いられていることを示している。しかしながら，必ずしも前提知識を完全に理解してから依存す

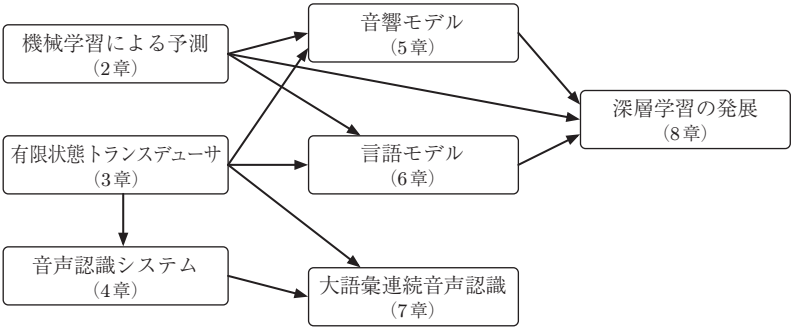

図 1.2 本書の構成

る章に進む必要はない。必要に応じて前の章に戻れるように，必要な前提知識が登場するたびに参照を付記する。

本書では，まず 2 章で機械学習の基本を解説する。つぎに 3 章で有限状態トランスデューサを解説する。これら 2 つの章は音声認識に至る前の，準備のための章である。これらのトピックについて基礎的な知識をすでに持っている読者は飛ばしてもよい。続けて，4 章では音声認識システムの大まかな構成を，5 章では音声信号と音素の対応を記述する音響モデルを，6 章では言語の構造を記述する言語モデルを解説する。また，7 章では，これまでの章で紹介した技術をまとめて，大語彙連続音声認識システムを実装する際の技術について解説する。最後に，8 章において，近年の発展が目ざましい深層学習技術の進展を音響モデル/言語モデルの区切りなく紹介する。

1.3　本書で用いる数式の表記

系 列 の 表 記

集合 S の要素からなる系列全体の集合を S^* と表記する。長さが他の変数によって表されるとき，例えば長さ T の実数系列であれば $\mathbb{R}^T$ のように表記する。この記述からわかるように，本書では長さ D の実数の系列の集合と，D 次元の実数ベクトルの集合を，$\mathbb{R}^D$ として同一視し，これらの要素について共通の演算を利用する。例えば，2 つのスカラー系列 $\boldsymbol{x}, \boldsymbol{y}$ について演算 $\boldsymbol{x}+\boldsymbol{y}$ は 2 つの系列の長さが同じときに定義される演算であり，各成分の和をとる演算とする。また，実数ベクトル系列（例えば $\mathbf{X} \in (\mathbb{R}^D)^T$）は実行列 $\mathbb{R}^{D\times T}$ と同一視される。例えば，行列 $\mathbf{A} \in \mathbb{R}^{D\times D'}$ をベクトル系列 $\mathbf{X} \in \mathbb{R}^{D'\times D''}$ に左から乗じて，新たな系列 $\mathbf{AX} \in \mathbb{R}^{D\times D''}$ を得ることができる。

系列変数 $\boldsymbol{x}$ の n 番目から m 番目までの要素を含む部分系列は，$\boldsymbol{x}_{n:m}$ と記述する。Python のスライス記法に似た表記であるが，終端，すなわち m 番目の要素を含むことと，断りのない限り，最初の要素の添字が 1 から始まる点に注意されたい。

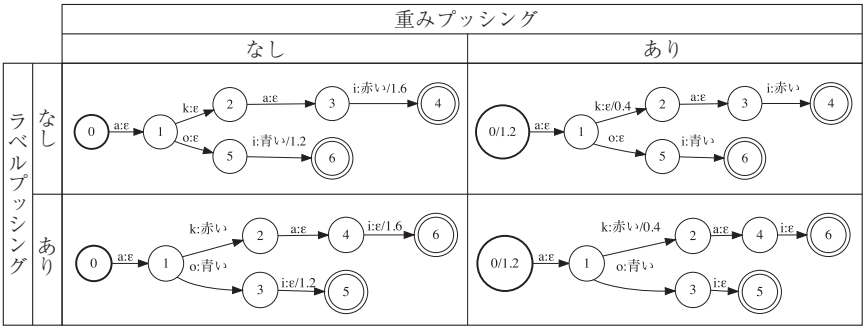

図 3.12 プッシング操作（重みをトロピカル半環として実行した場合）

形で等価性を保っている†。また，ラベルと重みの両方をプッシングすることもある（右下）。その場合は，両方の特性を持つ FST が得られる。いずれの場合も，プッシングは FST の等価性を保つ。

重みのプッシングアルゴリズムは，**アルゴリズム 3.7** のように表される。

アルゴリズム 3.7 重みプッシングアルゴリズム

入力 FST $\boldsymbol{X} = (Q, \Sigma, \Delta, I, F, E)$
出力 プッシングされた FST $\boldsymbol{Y}$
$E^{\mathtt{Rev}} \leftarrow \{(q', \sigma, \delta, q, w); (q, \sigma, \delta, q', w) \in E\}$
$\boldsymbol{X}^{\mathtt{Rev}} \leftarrow (Q, \Sigma, \Delta, F, I, E^{\mathtt{Rev}})$
// 積の交換法則を満たさない重みを用いる場合，ForwardScore 内の $\otimes$-積の順序を交換する必要がある点に注意
$D \leftarrow \textsc{ForwardScore}(\boldsymbol{X}^{\mathtt{Rev}})$ // → アルゴリズム 3.3
$I' \leftarrow \{(q, w \otimes D[q]) \colon (q, w) \in I\}$, $F' \leftarrow \{q, (D[q])^{-1} \otimes w \colon (q, w) \in F\}$
$E' \leftarrow \{(q, \sigma, \delta, q', (D[q])^{-1} \otimes (w \otimes D[q'])) \colon (q, \sigma, \delta, q', w) \in E\}$
$\boldsymbol{Y} = (Q, \Sigma, \Delta, I', F', E')$

このアルゴリズムでは，まず前向きスコア計算のアルゴリズムを，状態遷移を式 (3.30) に従って逆方向にした FST $\boldsymbol{X}^{\mathtt{Rev}}$ に適用することによって，各状態から終了状態に至るまでのパスの重みの $\oplus$ 総和，すなわち後向きスコアを得る。ある状態 q が後向きスコア $D[q]$ を持つということは，その状態から最終状態に至

† 初期状態重みをサポートしていないツールキットでは，初期状態重みは最初の状態遷移の重みとして付与される。

の重みを $\oplus$ 演算で合計したものを最小にする候補を探索しなければならないところを，重みを最小にする 1 つのパスを探索する問題に置き換えて近似していることになる。このような解の近似探索法を**ビタビデコーディング法**（Viterbi decoding）[16] と呼ぶ。ビタビデコーディングは近似的な求解法ではあるものの，多くの問題では実用上十分な精度を持つ。

結果として得られた $\hat{\boldsymbol{y}}$ を**図 5.4** に示す。学習データに対する評価においても，モデル化による誤差のため，データの完璧な再現ができていないことに着目しよう。本来，性能が過大評価されてしまう学習データに対する評価においても十分な性能が得られていないこの状況は，データに対してモデルの表現力が不足していることを示唆している。

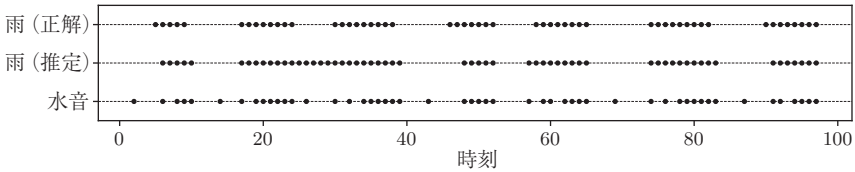

図 5.4　HMM による予測

5.1.2　複数の HMM 状態を持つモデル

状態数を増やしてモデルの表現力を向上させることで，学習データの再現性を高めることができる。生成モデルは，データをより良く表現する合理的なモデルを選択することによって精密化される。この例のデータの場合，データを細かく観察すると，例えば，水音の頻度は雨の降り始めと降り終わりで異なるという仮説を立てることができる。この仮説に従って，雨の前半と後半を個別の出力分布パラメータで表すことを考える。

雨の状態を 2 つに分けることを考え，HMM 状態変数 $s_t \in \{$ 晴$_1$, 雨$_1$, 雨$_2$$\}$ を導入する†。これまでの設定と異なり，天気の分布 $p(\boldsymbol{y})$ のみがマルコフ連鎖

† HMM の隠れ変数として用いられる状態変数と，FST の状態は，似て非なる概念なので注意が必要である。HMM の内部状態は，FST による表現では FST の状態遷移に付随する入出力アルファベットとして扱われる。本書では，「状態」という単語は基本的に FST の状態を表すこととし，HMM の隠れ変数として用いられる隠れ状態は「HMM 状態」の語で表す。

に従うのではなく，追加で導入した隠れ変数の確率分布 $p(\boldsymbol{s} \mid \boldsymbol{y})$ もマルコフ連鎖に従うと考える。HMM の基本的な定式化では，出力分布，すなわち観測変数の分布は，隠れ変数の実現値ごとに個別のパラメータを持つように設計される。したがって，HMM の隠れ変数が多くの値をとりうるように設計することで，モデルの表現能力を高めることができる。

この設定では，$y_t =$ 晴 のときの HMM 状態変数は，これまでどおり 1 つの値をとる。すなわち，$y_t =$ 晴 であれば，$p(s_t \mid y_t) = \mathbf{1}\,[s_t = \text{晴}_1]$ である。$y_t =$ 雨 となる区間では，それを表す出力分布を分けることを目的に，2 つの HMM 状態で前半と後半に分割する。$y_t =$ 雨 となる区間の中で，前半に対応する時刻 t では $s_t = \text{雨}_1$ となるように，また後半に対応する時刻 t では $s_t = \text{雨}_2$ となるように，HMM 状態の遷移を設計したい。**図 5.5** は，このような HMM 状態の動きを FST によって表現した例である。

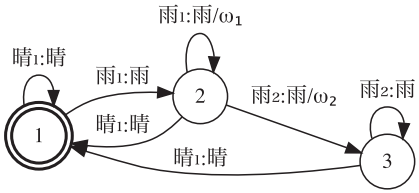

図 5.5　天気の状態遷移 FST（雨を 2 状態と見なした場合）

この FST が受理する入力の系列に注目する。この FST の動作は，晴$_1$ を何度か受理することから始まる。そして，晴から雨に変わるとき，すなわち雨に対応する HMM 状態を最初に受理するときに，雨$_1$ を受理する。その後，雨$_1$ を何度か繰り返し受理した後，$e^{-\omega_2}$ の確率で 雨$_2$ を受理する状態に入る。どちらの状態からも，晴$_1$ を受理して晴の天気に対応する状態に移行することもできる。

重みの具体的な推定法については後述するが，ここでは，重みは 1 階のマルコフ連鎖 $p(s_t \mid s_{t-1}, \boldsymbol{y})$ を表現するように設定されていることとする。状態 2 に注目すると，3 つの状態遷移が状態 2 から出発しているが，$y_t =$ 晴 を出力する状態遷移は 1 つだけであり，そのような状態遷移に対応する確率は 1 とな

とを考える。このようなパスは，メモリセルを辿るパスと，前時刻の出力を辿るパスの 2 通りがあり，特にメモリセルを辿るパスに活性化関数やアフィン変換が含まれないことが重要である。活性化関数は勾配のスケールを小さくするため，勾配消失の原因になり，また，アフィン変換はパラメータの大きさによって勾配を消失させたり爆発させたりする。LSTM は，離れた時刻の入力と出力との関連を，メモリセルを通じて勾配消失に悩まされることなく表現できる。

コーヒーブレイク

再帰結合ニューラルネットの双方向化

ここまでで見てきた再帰結合ニューラルネットの構成はすべて，過去の時刻の入力をなんらかの手段で記憶し，未来の時刻での予測に利用するものであった。しかし，未来の時刻における入力が，過去の時刻での予測に役立つ状況があることは想像に難くない。このような場合，RNN の時間的な前後関係を逆転した逆方向 RNN が用いられる。

前後両方の情報を考慮した予測を行うためには，**双方向 RNN**（bi-directional RNN）と呼ばれる構成がとられる [13)]。この構成では，順方向の RNN と逆方向の RNN を並列し，それらの結果を加算やベクトル連結によって統合した後に，つぎの層に渡す。この要素となる順方向/逆方向 RNN は，複数層からなってもよいし，双方向 RNN ブロック，すなわち順方向 RNN と逆方向 RNN を内部に持ち，これらの結果を統合するブロックを複数回繰り返してもよい。要素 RNN に LSTM を用いた**双方向 LSTM**（bi-directional LSTM）は，8 章で述べる End-to-end 音声認識の基本構成要素として広く用いられている。

双方向 RNN は，高い表現力を持つ重要な道具である。しかし，双方向 RNN は，逆方向の RNN の最初の入力として，入力系列の最後の要素を必要とするため，入力系列をすべて受け取ってからでないと計算を始められないという欠点がある。実システムでは，しばしばオンライン性，すなわち入力信号を受け取りながら認識を行える性質が重要であり，双方向 RNN にオンライン性を与えるには，さまざまな工夫が必要である。

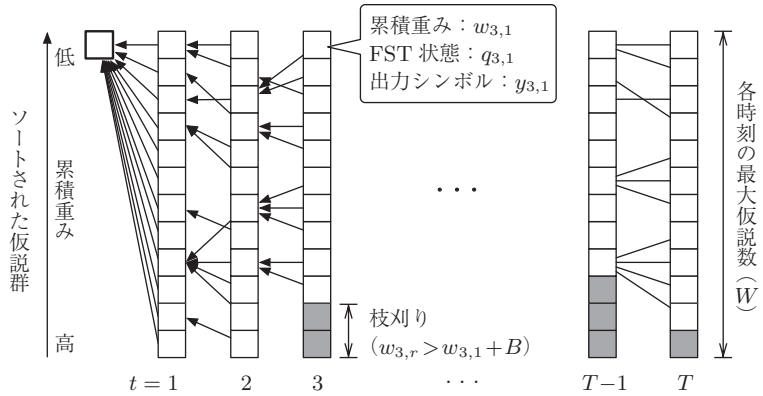

図 7.3　時刻同期ビーム探索のデータ構造

図 7.3 に，ビーム探索に利用されるデータ構造の一例を示す†。

トレリス FST の状態変数は $Q[\boldsymbol{E}] \times Q[\boldsymbol{D}] = \{0, \cdots, T\} \times Q[\boldsymbol{D}]$ であり，各状態は，時刻 t とデコーディングネットワーク上の状態変数 $q \in Q[\boldsymbol{D}]$ のペア $(t, q) \in Q[\boldsymbol{E} \circ \boldsymbol{D}]$ で表現される。時刻同期探索では，トレリスの FST の状態を時刻 t ごとにグループ化する。本来の最短経路の探索問題では，初期状態からある状態 (t, q) に至るすべての経路の重みの $\oplus$ 和である前向きスコアを計算する（3.5 節参照）が，規模の大きさからそれが不可能な場合の近似として，グループごとに，そこに至るまでの重みが最も小さいいくつかの状態のみを利用し，他の状態を辿る経路を無視（枝刈り）して前向きスコアを計算していく手法がとられる。これをビーム探索と呼ぶ。

時刻 t ごとにデコーディングネットワークの状態 q，前向きスコア w'，出力シンボル列 $\boldsymbol{y}$ とバックポインタ b の 4 つ組 $(q, w', \boldsymbol{y}, b) \in (Q[\boldsymbol{D}] \times \mathbb{K} \times \Delta^* \times \mathbb{N})$ の配列 $H[t]$ が，w' が小さい順にソートされているとする。バックポインタ b は，時刻 $t-1$ にデコーディングネットワークのどの状態に至っていたのかを，

† 音声認識の探索アルゴリズムは，音声認識利用時の計算コストに最も影響する部分であり，高速計算のテクニックや近似が多用される。加えて，音声認識の結果を出力する部分であるため，その応用において求められる機能に応じて拡張されることも多い。本節で解説するのはその最も簡素な実現法であり，現実のシステムにおける複雑なデコーダの実装と乖離していることに注意されたい。

索　　引

—— 著 者 略 歴 ——

久保　陽太郎（くぼ　ようたろう）

2007年　早稲田大学理工学部情報学科卒業
2008年　早稲田大学大学院基幹理工学研究科修士課程修了（情報理工学専攻）
2010年　早稲田大学大学院基幹理工学研究科博士課程修了（情報理工学専攻），博士（工学）
2010年　RWTH アーヘン工科大学客員研究員
2010年　日本電信電話株式会社コミュニケーション科学基礎研究所
2014年　Amazon，Speech Scientist
2019年　Google，Research Scientist
　　現在に至る

機械学習による音声認識

Machine Learning in Automatic Speech Recognition

2021 年 5 月 6 日　初版第 1 刷発行

検印省略

編　　者	一般社団法人 日本音響学会
発 行 者	株式会社　コ ロ ナ 社
	代 表 者　牛 来 真 也
印 刷 所	三 美 印 刷 株 式 会 社
製 本 所	牧 製 本 印 刷 株 式 会 社

112–0011　東京都文京区千石 4–46–10
発 行 所　株式会社　コ ロ ナ 社
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ　https://www.coronasha.co.jp

ISBN 978–4–339–01139–5　C3355　Printed in Japan　（新宅）G