

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

音響テクノロジーシリーズ編集委員会

編集委員長

千葉工業大学

博士（工学） 飯田 一博

編集委員

東北学院大学

博士（情報科学） 岩谷 幸雄

甲南大学

博士（情報科学） 北村 達也

滋賀県立大学

博士（工学） 坂本 眞一

国立音楽大学

博士（工学） 三浦 雅展

千葉工業大学

博士（工学） 大川 茂樹

東京大学

博士（工学） 坂本 慎一

神戸大学

博士（工学） 佐藤 逸人

（五十音順）

（2021年3月現在）

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る。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、スマートスピーカー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つであると考えられることもできる。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、荻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

1.1 本書の目的	1
1.2 本書の構成	3
1.3 本書で用いる数式の表記	4
1.4 確率論の基礎	7
1.4.1 周辺化	8
1.4.2 条件付き確率	9
1.4.3 独立性	10
1.4.4 連続分布と確率密度関数	11

2. 機械学習による予測

2.1 モデルによる予測	13
2.2 識別関数の構成	14
2.3 確率モデルの学習	16
2.4 最適化のアルゴリズム	20
2.4.1 凸関数の最適化	20
2.4.2 指数型分布族の最尤推定	25
2.4.3 潜在変数モデルとEMアルゴリズム	30
2.4.4 勾配に基づく局所最適化	38
2.5 例：身長と体重から学年を推定する	41
2.5.1 生成モデルによるアプローチ	42

2.5.2	識別モデルによるアプローチ	45
2.5.3	識別関数法によるアプローチ	48
2.6	深層学習	50
2.6.1	識別モデルの構成とソフトマックス層	53
2.6.2	確率的勾配降下法	54
2.7	モデル選択と過学習	58
2.7.1	過学習	59
2.7.2	交差検証	61
2.7.3	正則化	62
2.7.4	アーリーストッピング	63
	引用・参考文献	64

3. 有限状態トランスデューサ

3.1	有限状態オートマトン	65
3.2	文法と辞書の表現	68
3.2.1	重みの導入	69
3.2.2	トランスデューサの導入	70
3.3	有限状態トランスデューサの数学的定義	72
3.3.1	半環	72
3.3.2	状態集合 Q と状態遷移集合 E	74
3.3.3	初期状態 I と終了状態 F	75
3.3.4	遷移パスと重み	76
3.3.5	FST の等価性	78
3.3.6	対数確率半環と FST の確率的解釈	78
3.3.7	FST の連結, クリーネ閉包, 和	80
3.4	合成	82
3.4.1	合成演算のアルゴリズム	82
3.4.2	合成演算の確率的解釈	87
3.4.3	アルファベット列の FST による表現と合成演算	88
3.5	最短経路問題	89

3.6 FST の最適化	92
3.6.1 トリミング	92
3.6.2 ϵ 除去	94
3.6.3 重みとラベルのプッシング	98
3.6.4 決定化	104
3.6.5 最小化	111
3.7 対数確率半環の重みを持つ非巡回 FST 上の期待値計算	115
3.7.1 非巡回 FSA のトポロジカルソート	116
3.7.2 期待値計算	117
引用・参考文献	122

4. 音声認識システム

4.1 音声認識システムの構成	123
4.2 音声の単位	125
4.2.1 音素を介した音声認識の生成モデル	127
4.2.2 発音辞書モデル	128
4.3 音声の分析	130
4.3.1 音声信号のモデル	131
4.3.2 離散フーリエ変換と周波数解析	132
4.3.3 フィルタバンク処理	138
4.3.4 ケプストラム抽出と無相関化	142
4.3.5 対数エネルギー	143
4.3.6 セグメント分析	143
4.4 音声認識システムの評価法	146
4.4.1 認識精度の評価	146
4.4.2 計算効率の評価	150
引用・参考文献	151

5. 音響モデル

5.1 隠れマルコフモデル	152
5.1.1 雨と水音のモデル	153
5.1.2 複数の HMM 状態を持つモデル	158
5.1.3 雨の推定から音声認識へ	170
5.2 混合正規分布と連続分布型 HMM	173
5.3 音素文脈依存モデル	178
5.3.1 決定木による音素文脈クラスタリング	179
5.3.2 決定木を用いた音響モデルの FST による表現	187
5.3.3 凝集型クラスタリングによる質問の自動生成	189
5.4 ニューラルネットによる音響モデル	192
5.4.1 再帰結合ニューラルネット	194
5.4.2 ゲートユニットと長短期記憶	196
5.5 系列識別学習	200
5.5.1 系列識別学習規準	200
5.5.2 認識仮説を用いた最適化アルゴリズム	205
5.6 音響モデル適応の技術	208
5.6.1 声道長正規化による適応	209
5.6.2 話者コードの入力による適応	211
5.6.3 再学習による適応	212
引用・参考文献	212

6. 言語モデル

6.1 言語モデルとは	215
6.2 ユニグラム言語モデルと Bag-of-words	218
6.3 N グラム言語モデル	220



6.4	N グラム言語モデルの学習と平滑化	221
6.4.1	N グラム言語モデルの最尤推定	222
6.4.2	加算平滑化	224
6.4.3	線形補間平滑化	226
6.4.4	ウィトソン・ベル平滑化	227
6.4.5	グッド・チューリング推定法	228
6.4.6	カット平滑化	231
6.4.7	絶対割引法	233
6.4.8	クニーザー・ナイ平滑化	234
6.5	N グラム言語モデルの FST による表現	235
6.6	最大エントロピーモデルと識別的言語モデル	238
6.6.1	最大エントロピー原理に基づく言語モデル	238
6.6.2	文レベルの最大エントロピーモデル	244
6.6.3	音声認識のための識別的言語モデル	245
6.7	ニューラルネット言語モデル	247
6.7.1	ニューラルネットによる後続単語の予測	248
6.7.2	単語の分散表現	251
6.7.3	ニューラルネット言語モデルによるリスクアリング	252
	引用・参考文献	254

7. 大語彙連続音声認識

7.1	FST の合成と確率モデル	256
7.1.1	デコーディングネットワークの構成と探索誤り	257
7.1.2	非曖昧化シンボル	258
7.2	大語彙連続音声認識の探索問題	261
7.3	大規模 FST 合成の技術	266
7.3.1	オンザフライ合成	266
7.3.2	データベース認識システム	270
7.4	N ベストリストおよびラティスの生成	271
7.4.1	ラティスの生成	271

7.4.2 ラティスからの N ベストリストの生成 273

引用・参考文献.....275

 **8. 深層学習の発展** 

8.1 さまざまなニューラルネット要素.....276

8.1.1 飽和しない活性化関数 276

8.1.2 ドロップアウト 278

8.1.3 バッチ正規化 281

8.1.4 畳み込み層/プーリング層 284

8.2 ニューラルネットの高速化.....287

8.2.1 重みの量子化 287

8.2.2 特異値分解による重み行列の圧縮 288

8.2.3 蒸留によるモデル変換 290

8.3 End-to-end 音声認識.....292

8.3.1 CTC 292

8.3.2 エンコーダ-デコーダ型 End-to-end 音声認識 295

引用・参考文献.....303

索引.....306



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

2 1. 本書の目的と事前知識

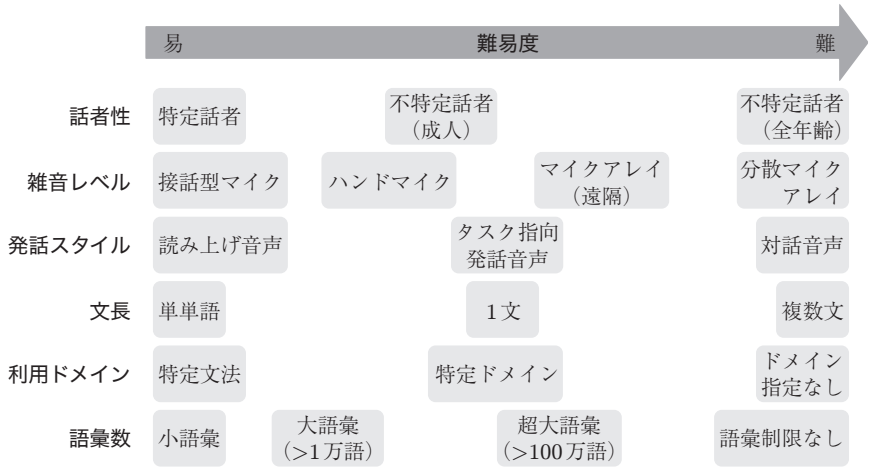


図 1.1 音声認識の達成目標と難易度

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「入力に奇数のときに 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

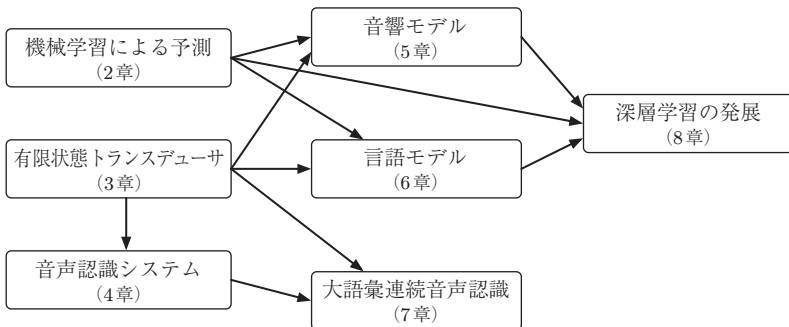


図 1.2 本書の構成

4 1. 本書の目的と事前知識

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 \mathbf{x}, \mathbf{y} について演算 $\mathbf{x} + \mathbf{y}$ 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 \mathbf{x} の n 番目から m 番目までの要素を含む部分系列は、 $\mathbf{x}_{n:m}$ 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。

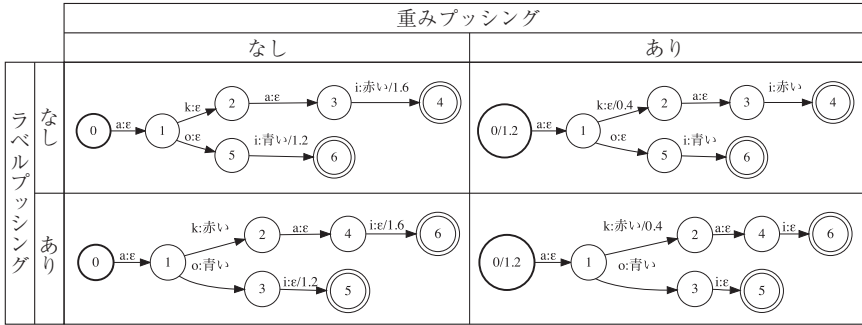


図 3.12 プッシング操作 (重みをトロピカル半環として実行した場合)

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  $X = (Q, \Sigma, \Delta, I, F, E)$ 
出力 プッシングされた FST  $Y$ 
 $E^{\text{Rev}} \leftarrow \{(q', \sigma, \delta, q, w); (q, \sigma, \delta, q', w) \in E\}$ 
 $X^{\text{Rev}} \leftarrow (Q, \Sigma, \Delta, F, I, E^{\text{Rev}})$ 
// 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
 $D \leftarrow \text{FORWARDSCORE}(X^{\text{Rev}})$  //  $\rightarrow$  アルゴリズム 3.3
 $I' \leftarrow \{(q, w \otimes D[q]): (q, w) \in I\}, F' \leftarrow \{q, (D[q])^{-1} \otimes w: (q, w) \in F\}$ 
 $E' \leftarrow \{(q, \sigma, \delta, q', (D[q])^{-1} \otimes (w \otimes D[q'])): (q, \sigma, \delta, q', w) \in E\}$ 
 $Y = (Q, \Sigma, \Delta, I', F', E')$ 
    
```

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 X^{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することになる。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{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。

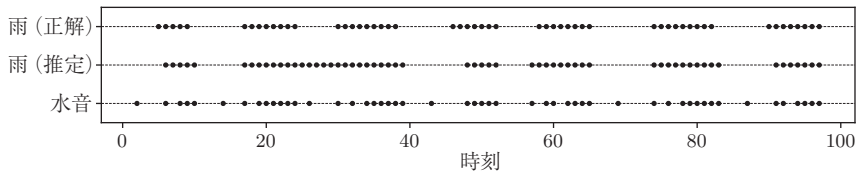


図 5.4 HMM による予測

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 $s_t \in \{\text{晴}_1, \text{雨}_1, \text{雨}_2\}$ を導入する[†]。これまでの設定と異なり、天気分布 $p(y)$ 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 $p(\mathbf{s} | \mathbf{y})$ 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、 $y_t = \text{晴}$ 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、 $y_t = \text{晴}$ であれば、 $p(s_t | y_t) = \mathbf{1}[s_t = \text{晴}_1]$ である。 $y_t = \text{雨}$ 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。 $y_t = \text{雨}$ 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。

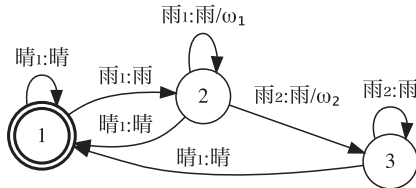


図 5.5 天気の状態遷移 FST (雨を 2 状態と見なした場合)

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 $p(s_t | s_{t-1}, \mathbf{y})$ 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、 $y_t = \text{晴}$ 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、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。

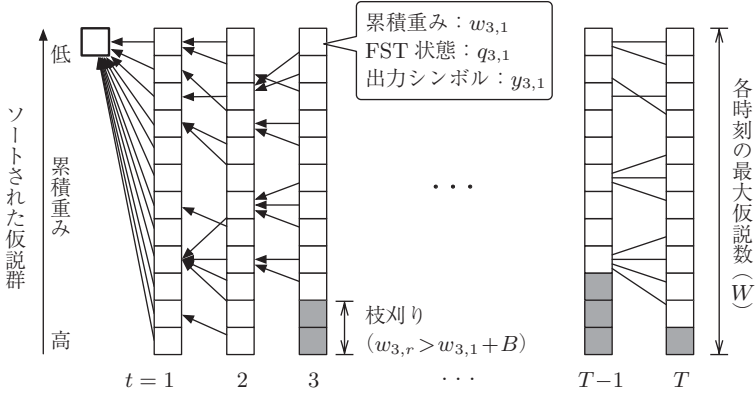


図 7.3 時刻同期ビーム探索のデータ構造

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 $Q[\mathbf{E}] \times Q[\mathbf{D}] = \{0, \dots, T\} \times Q[\mathbf{D}]$ 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 $q \in Q[\mathbf{D}]$ のペア $(t, q) \in Q[\mathbf{E} \circ \mathbf{D}]$ 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 \mathbf{y} とバックポインタ b の 4 つ組 $(q, w', \mathbf{y}, b) \in (Q[\mathbf{D}] \times \mathbb{K} \times \Delta^* \times \mathbb{N})$ 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

【あ】	
アーリーストッピング法	37, 63, 182
アダマール積	137, 197
アテンション機構	298
アフィン変換	51
アンサンブル学習	278
【い】	
イェンセンの不等式	21, 163
異音	125, 178
【う】	
ウイトン・ベル平滑化	228
ウォームスタート	63
【え】	
枝刈り	261, 263, 300
エラー削減率	149
エンコーダ	296
エントロピー	202, 240
【お】	
オイラーの公式	133
凹関数	21
重み付き有限状態オートマトン	70
重み付き有限状態トランスデューサ	65, 70
音響特徴ベクトル	131
音響特徴量	131

音響モデル	152
音響モデルスケール	172, 246
オンザフライ合成	267
音素	125
音素文脈依存モデル	179

【か】

開発データセット	36, 61
過学習	59
学習グラフ	167
確率質量関数	8
確率的勾配降下法	55
確率変数	12
確率密度関数	11
隠れ変数	30, 153
隠れマルコフモデル	152
加算平滑化	224
活性化関数	51
カット平滑化	231
カットオフ	238
カテゴリカル分布	27
ガリック半環	101
カルバック・ライブラー情報量	32, 291
関数型	108
慣性項	58
ガンマ関数	225

【き】

機械学習	1, 13
期待値最大化法	31, 162
逆離散フーリエ変換	133

教師あり学習	14
共分散行列	40, 174
共変量シフト	281

【く】

グッド・チューリング推定法	228
クニエザー・ナイ平滑化	234
グラフ走査	93
クリーネプラス	81
クリーネ閉包	81
クロスエントロピー	53
クロネッカーのデルタ関数	37

【け】

経験誤差	58
形態素	217
形態素解析器	217
系列アラインメント	170
系列最小ベイズリスク	203
決定化	104
決定木	179
決定木クラスタリング	182
決定性オートマトン	67
ケプストラム	142
ケプストラム分散正規化	145
ケプストラム平均正規化	145
言語モデルスケール	173

【こ】

高域強調	134
交互最適化	32

交差検証 61
 合成アルゴリズム 82, 266
 構造識別器 14
 高速フーリエ変換 136
 勾配降下法 39
 勾配消失 196, 276
 国際音声記号 125
 混合正規分布 30, 176
 混合分布 30, 176, 226

【さ】

再帰結合ニューラルネット 194, 249
 最急降下法 39
 最小誤り率学習 246
 最小音素誤り 203
 最小化 111
 最小単語誤り 203
 最大エントロピー原理 240
 最大エントロピーマルコフモデル 239
 最大相互情報量推定 200
 最短経路問題 89, 261
 最長共通接頭辞 101
 再統合 263
 最尤推定 19, 155, 162, 221
 最尤線形変換 144
 最良優先探索 273
 削除誤り 147
 サブサンプリング 286
 サブワード 217
 サポートベクターマシン 48

【し】

ジェリネック・マーサー平滑化法 226
 時間遅れニューラルネット 284
 識別関数法 16, 48
 識別モデル 16, 45
 自己回帰型ニューラルネット 296
 事後確率 10

時刻同期ビーム探索 261
 自己符号化器 279
 事後分布 10
 事象 7
 指数移動平均法 56
 指数型分布族 25
 事前確率 10
 自然パラメータ 26
 事前分布 10
 十分統計量 28, 174
 周辺化 8
 終了状態 66
 出力分布 155
 条件付き最尤推定 19, 201
 条件付き独立性 154
 条件付きモデル 16
 状態占有率 169
 情報量 202, 240
 蒸留 290
 初期状態 66
 深層学習 50, 276
 深層ニューラルネット 52, 193

【す】

数理モデル 13
 スキップグラム 251
 スペクトル 133
 スペクトル包絡 138
 スラック変数 48

【せ】

正規表現 65
 正規分布 16, 27, 173
 制限ボルツマンマシン 279
 成功パス 76
 生成モデル 16, 42
 正則化 62
 正則化項 49
 正定値行列 174
 声道長正規化 209
 制約付き最適化 22
 積の法則 10

絶対割引法 233
 接頭辞 105
 接尾辞 111
 セルフアテンション 300
 ゼログラム 226
 ゼロ頻度問題 221
 線形識別分析 144
 線形時不変システム 131
 線形補間 226
 潜在変数 30

【そ】

総合重み 77
 相互情報量 201, 202
 挿入誤り 147
 双方向 LSTM 199
 双方向 RNN 199
 素性関数 239
 素性ベクトル 239
 ソフトマックス層 53
 損失関数 54

【た】

大語彙連続音声認識 256
 対数確率半環 73
 対数線形 242
 対数分配関数 28
 対数メルフィルタバンク出力 141
 多クラス識別器 14
 畳み込み 131, 284
 畳み込み層 284
 畳み込みニューラルネット 287
 多変量混合正規分布 170
 多変量正規分布 27, 173
 単音 125
 単語誤り率 147
 単語同期ビーム探索 300
 単語分割 217
 探索誤り 257
 短時間フーリエ変換 134

【ち】		【は】		分配関数	28
遅延評価	267	パープレキシティ	219	文脈非依存モデル	180
置換誤り	147	バイト対符号化	217	【へ】	
調音結合	178	ハイパーパラメータ	61	ベイキス型 HMM	171
長短期記憶	197	ハイブリッド型	192	平均ベクトル	174
直積半環	73	バウム・ウェルチ法	166	バイズ推論	17
【つ】		白色化	40	バイズ則	10
通時的誤差逆伝播法	195	発音辞書モデル	128	ベースラインシステム	149
【て】		バックオフ	233	ベルヌーイ分布	24, 27
ティホノフの正則化	63	幅優先探索	261	【ほ】	
テイラー展開	38	ハミング窓	136	補 間	232
ディリクレ分布	27, 225	パラメータ	16	ボトム	85
データ増強法	210	汎化誤差	58	【ま】	
デコーダ	295	半 環	72	マージン	48
点推定	18	万能性定理	52	前向き後向きアルゴリズム	121, 166
【と】		【ひ】		窓関数	135
統計的言語モデル	216	非曖昧	157	マルコフ連鎖	154
動的計画法	118	非曖昧化シンボル	259	マルチヘッドアテンション	300
トークン化	217	ビーム探索	261, 300	【み】	
特異値分解	288	非決定性オートマトン	67	ミニバッチ	55
特徴抽出	130	非巡回	115	ミニマルペア	125
独 立	11	ヒストグラムプルーニング法	264	【め】	
独立同分布	17	ビタビアラインメント	170	メル尺度	138
凸関数	20	ビタビデコーディング法	158, 257	メル周波数ケプストラム	142
トポロジカルソート	116	評価データセット	61	係数	142
トライフォン	179	【ふ】		メルフィルタバンク	138
トランスフォーマー	300	フィードフォワード型		【も】	
トリミング	92	ニューラルネット	50	モーメンタム項	58
トレリス	157	フィルタバンク	139	文字列半環	73
ドロップアウト法	278	ブーステッド最大相互		モデル適応	208
トロピカル半環	73	情報量推定	201	モンテカルロ法	54
【な】		プーリング層	284	【ゆ】	
ナイキスト周波数	140	プール半環	73	有限状態アクセプタ	66
内部共変量シフト	282	深さ優先探索	93	有限状態オートマトン	65
【に】		プッシング	98		
ニューラルネット	50, 192, 247, 276	フレーム	134		
		文誤り率	146		
		分散表現	251		

有限状態トランスデューサ	ラティス	206, 271
2, 65, 70	ラプラス平滑化	224
連結	【り】	
和	リアルタイムファクタ	150
尤度	離散コサイン変換	142
ユニグラム言語モデル	離散フーリエ変換	132
ユニタリ行列	リスコアリング	243, 253
【ら】	量子化	287
ラグランジュ関数	【る】	
ラグランジュの未定乗数	ルックアヘッド合成	267
ラグランジュの未定乗数法		
22		

【れ】	
レーベンシュタイン距離	147
レキシコン	128
【ろ】	
ロジスティック回帰	45
【わ】	
和の法則	9
割引	224

【A】	
ARPAbet	126
【B】	
bag-of-words	218
bMMI	
⇒ブーステッド最大相互情報量推定	
BPTT ⇒通時的誤差逆伝播法	
【C】	
CBoW	251
CTC	293
【D】	
DCT ⇒離散コサイン変換	
DNN ⇒深層ニューラルネット	
【E】	
EM アルゴリズム	
⇒期待値最大化法	
end-to-end 音声認識	292
【F】	
FSA ⇒有限状態アクセプタ	

FST	
⇒有限状態トランスデューサ	
【G】	
GCLM	246
GMM ⇒混合正規分布	
【H】	
HMM ⇒隠れマルコフモデル	
【I】	
iid ⇒独立同分布	
IPA ⇒国際音声記号	
【K】	
k 平均法	38
KL ダイバージェンス	
⇒カルバック・ライブラー情報量	
【L】	
Left-to-right 型 HMM	171
LSTM ⇒長短期記憶	
LVCSR	
⇒大語彙連続音声認識	

【M】	
MAP 推定	18
MEMM	
⇒最大エントロピーマルコフモデル	
MERT ⇒最小誤り率学習	
MFCC	
⇒メル周波数ケプストラム係数	
MLE ⇒最尤推定	
MMI 推定	
⇒最大相互情報量推定	
MPE ⇒最小音素誤り	
MWE ⇒最小単語誤り	
【N】	
N グラム	223
N グラム言語モデル	220
N ベストリスト	206
【O】	
One-hot 表現	248
【P】	
pdf ⇒確率密度関数	

	[R]	SVD ⇒特異値分解	WFST
RBM		SVM ⇒サポートベクターマシン	⇒重み付き有限状態トラン スデューサ
⇒制限ボルツマンマシン		[T]	word2vec 251
RNN		TDNN	[X]
⇒再帰結合ニューラルネット		⇒時間遅れニューラルネット	X-SAMPA 126
RTF		teacher forcing 297	【数字・記号】
⇒リアルタイムファクタ		[V]	0-1 損失関数 58
[S]		VTLN ⇒声道長正規化	⊥ 85
SGD ⇒確率的勾配降下法		[W]	ε シーケンシングフィルタ 86
sMBR		weakly left-divisible 106	ε 除去アルゴリズム 94
⇒系列最小ベイズリスク			ε 閉包 94
STFT			
⇒短時間フーリエ変換			

—— 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

© 一般社団法人 日本音響学会 2021

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



本書のコピー、スキャン、デジタル化等の無断複製・転載は著作権法上での例外を除き禁じられています。購入者以外の第三者による本書の電子データ化及び電子書籍化は、いかなる場合も認めていません。落丁・乱丁はお取替えいたします。