# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 音声認識における”隠れマルコフモデル”と逆問題 |
| Title(English) | |
| 著者(和文) | 古井貞熙 |
| Authors(English) | SADAOKI FURUI |
| 出典(和文) | 統計数理研究所共同研究リポート, Vol. , No. 138, pp. 55-59 |
| Citation(English) | , Vol. , No. 138, pp. 55-59 |
| 発行日 / Pub. date | 2001, |

統計数理研究所共同研究リポート 138
逆問題とその周辺 (7)
(2001 年 3 月、統計数理研究所)

# 音声認識における "隠れマルコフモデル "と逆問題

古井　貞熙
東京工業大学大学院情報理工学研究科計算工学専攻
furui@cs.titech.ac.jp

## 1. はじめに

最近，音声認識技術は実用化に近いレベルに達しており，大語彙連続音声認識の技術も大きく進歩している[1][2]．このような背景のもとに，音声入力ワープロ，カーナビ，種々の情報案内への問い合わせなど，いろいろな音声認識応用が注目されるようになってきた．音声認識を用いてニュース音声を自動的に字幕化するシステムも開発されている．これらの背景には，HMM（隠れマルコフモデル; hidden Markov model）や統計的言語モデルに代表される音声認識技術の進歩とともに，その実現を可能にしたコンピュータやLSIの能力の急速な向上がある．

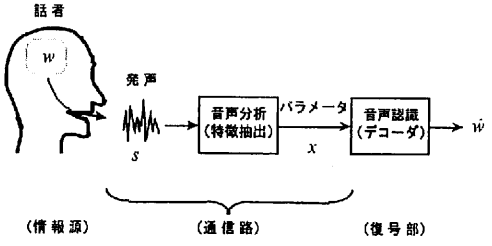
音声認識は，極めて多様性の大きい音声波形から，直接は観測できない，話者が何を発声した(発声しようとした）のかという情報，具体的には単語列あるいは話者の意図を推定する問題であり，一種の逆問題ということができる．ここでは，音声認識の基本技術として，HMM の利用法について解説する．

## 2. 音声認識の基本技術

### 2.1 音声認識の原理

音声認識は，図１に示すように，情報源，通信路，復号部からなるシステムに対する情報処理論・通信理論の問題として，確率モデルを用いて定式化することができる[3][4]．すなわち，話者（情報源）が頭の中で考えた内容（通常は単語列）を $w$ で表すと，$w$ は話者の発声器官によって音声波形 $s$ に変換される．音声認識システムでは，これをマイクロホンで電気信号に変換した後，波形の数値列としてコンピュータに取り込む．この数値列は，周波数スペクトルの時間変化を表す特徴パラメータ $x$ に変換される．特徴パラメータとしては，通常，対数スペクトルを逆フーリエ変換したケプストラムと，その１次および２次の回帰係数が用いられる．

この特徴パラメータ列から，元の話者の頭の中にあった単語列を推定（復号化）する．この際，パターン認識の理論から，事後確率を最大にする $w$ を選べば誤認識を平均的に最小にすることができるので，図の式に従って $w$ を決定する．通常，事後確率を直接計算することは困難なので，$w$ は，ベイズ則によって展開した右辺の式を最大化するように推定される．ここで，条件つき確率 $P(x|w)$は音素などの音響モデルから特徴パラメータ列が出現する確率として計算され，単語列 $w$ が発声される事前確率 $P(w)$は言語モデル（文法）によって計算される．




$$\hat{w} = \underset{w}{\operatorname{argmax}}\, P(w|x) = \underset{w}{\operatorname{argmax}}\, P(x|w)P(w)$$

図１　確率モデルに基づく音声認識システムの構成

人が音声を聞いて理解しようとするとき，音と

しての情報だけでなく，語彙に関する情報を始め，文法，意味など種々の情報を組み合わせて音声を聞き取っている．多くの場合，状況や話題の展開から無意識に単語や句を予測したり，文脈から推し量って認識している．その一つの理由は，音素情報のあいまい性である．すなわち，通常の音声の中では各音素は必ずしもはっきりと発声されず，その部分だけを聞いたのでは何の音素か分からないことがめずらしくない．このため，コンピュータによる音声認識においても，単語列の事前確率 $P(w)$の果たす役割が極めて大きい．言語情報のモデル化については，ここではこれ以上触れない．

### 2.2 HMM による音響モデルと音響処理

条件つき確率 $P(x|w)$を求めるためには，単語や音素などの音声の基本的な単位のモデルを蓄えておき，仮定した単語列 $w$ に従って，その単位を接続したものと，認識すべき音声を音響的に照合する．調音結合，すなわち各音素の物理的特性が前後の音素の影響によって変形する現象があるため，単語を単位とした方が，その変形を自然に取り込めるが，認識対象語彙数が大きくなると，記憶容量と認識のための計算量が膨大になってしまう．このため語彙数が大きいときには，音素を基本単位として用いて，前後の音素に依存した複数のモデルを用意することにより，調音結合に対処する．

音声の特徴パラメータ系列と，各音素や単語の基本単位との類似の度合いを計算する際に重要な課題の一つは，音声の時間的な伸縮にどう対処するかである．例えば，同じ音素でも，それが含まれる言葉，話者，発声速度などによって長さが変化する．しかも，その変化はゴム紐のように線形（一様）ではなく，部分的に伸びたり縮んだりする非線形な伸縮である．このようなパターンを，そのずれを調整しながら照合する方法として，動的計画法（DP; dynamic programming）を用いた方法（DP マッチングと呼ばれる）が広く用いられている[3].

音声には，時間軸だけでなく，スペクトルの形そのものが，個人差を始めとする多様な要因によって変化する性質がある．この問題に対処するため，図２に示すような HMM が広く用いられている[3][4]. HMM は，確率状態遷移機械であり，各状態遷移確率と，状態遷移に伴う特徴パラメータの観測（出現）確率によって特徴付けられる．これにより，音声の確率的変動が極めて効率よく表現できる．音声の各基本単位に対応する HMM をあらかじめ学習しておき，入力音声がその HMM の接続から生成される確率 $P(x|w)$を計算する．この計算も，DP を用いることによって能率よく行なうことができる．

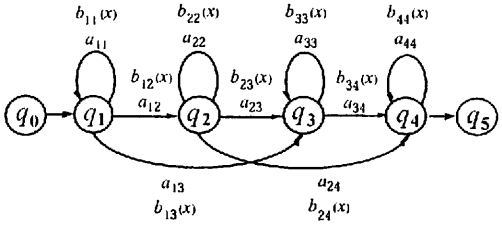



図２ 隠れマルコフモデル（HMM）の構成；$q_i$：状態，$a_{ij}$：遷移確率，$b_{ij}(x)$：スペクトルパターン $x$ の出現確率

### 2.3 出現確率分布の表現法

出現確率分布に関しては，連続値として表す場合（連続分布モデル，連続 HMM）と，有限個の符号ベクトルの組み合わせで表現する場合（離散分布モデル，離散 HMM）がある．連続分布モデルで出現確率を表す方法としては，単一ガウス分布（正規分布）または混合ガウス分布が用いられ，パラメータの自由度を減らすために無相関ガウス分布を用いることが多い．離散分布モデルの場合は，スペクトルパターンのクラスタ化（ベクトル量子化）によって，代表スペクトルパターン（符

号ベクトル）を生成し，各符号ベクトルの出現確率の組み合わせによって出現確率を表す．

連続分布モデルと離散モデルの中間として，半連続分布モデル（半連続 HMM）がある．これは，連続分布モデルにおける混合ガウス分布を，すべてのモデルのすべての状態で共通にし，各分布の重みだけを変えるようにしたものである．結び混合分布モデル（tied-mixture model）とも呼ばれる．離散分布モデルにおける各符号ベクトルに確率分布を持たせたものということもできる．

実際の音声認識に用いる HMM においては，対象に応じて適切に状態数やモデル構造（遷移構造）を決定し，スペクトルパターンの表現法（離散分布モデルの場合はその種類，連続分布モデルの場合はそのモデル化の方法）を決定する必要がある．これらの数や複雑さ，すなわちモデルの自由度を大きくすれば，きめ細かい変動が表現できるが，推定すべきモデルパラメータが多くなり，推定精度が悪くなる．

### 2.4 HMM の推定

HMM のパラメータは，大量の音声データ（学習用サンプル）を用いて，バウムウェルチ（Baum-Welch）アルゴリズムによって，尤度最大化規準で自動的に推定することができる．このアルゴリズムは，フォワードバックワード（forward-backward）アルゴリズムとも呼ばれ，期待値の計算と尤度最大化の計算が交互に繰り返される EM（expectation-maximization）アルゴリズムの一種である．少なくとも局所的最大値に収束することがわかっている．

## 3. 音声認識のための HMM の高度化

### 3.1 ロバストな音声認識を目指して

音声認識の難しさ，すなわち誤認識の主たる原因は，音響モデルや言語モデルの学習に用いたデータと，実際の認識すべき音声の間に，必ず何らかのずれ（mismatch）があることにある[2][3]．その背景には，音声（声質や話し方）の個人差（方言や年令によるものも含む）が極めて大きいこと，同じ人でも体調や精神状態（ストレス）などによって発声速度や話し方が変化すること，多くの場合に，部屋の反響を含む種々の雑音が音声に加わっていて，しかもそれが時間的に変化すること，さらにその上にマイクロホンや電話機，伝送系などの歪みが加わることなどがある．これらの変動に対して認識性能が低下しない音声認識法，すなわちロバスト（robust）な音声認識法を構築することが，極めて重要である．

### 3.2 音声変動への適応化

上記の種々の変動に対処するためには，音声認識システムに，少量の音声サンプルに基づいて，変動に自動的に追随してくれるような，自動適応機能を持たせることが必要である．一般に適応化アルゴリズムには，その規則を学習するためにあらかじめ決められた言葉を発声してもらう「教師つき（supervised）学習」と，任意の音声を用いて規則を学習する「教師なし（unsupervised）学習」による方法がある．前者の方が規則の獲得は比較的容易であるが，ユーザにとって，認識装置を使う前にわざわざ特定の言葉を発声しなければならないのは煩わしい．後者の場合は，認識すべき音声をそのまま使って適応化できるので，ユーザにとっては全く意識せずに，あたかも誰の声にも動作する装置であるように使えて便利である．さらに，認識装置を使っているうちに，自動的に適応化が進んでいくこと，すなわちオンライン逐次適応が望ましい．

音声は通常，スペクトルあるいは対数スペクトル領域のパラメータで表現されるが，上で述べた種々の音声の変動は，これらの領域での線形あるいは非線形の変化として表される．さらに，その変化が，元の音声パラメータに対して，加算的な

場合と乗算的な場合がある．この違いによって，適応化技術も異なってくる．また，これらの種々の変動は，単独にではなく複合して音声に加わるのが普通である．従って，これらに同時に対応できる適応化法を構築することが必要である．

### 3.3 主な適応化技術

特徴パラメータ領域の代表的な適応化法に，ケプストラムの時間平均値を差し引く方法（CMS; cepstral mean subtraction）がある．文の長さ程度にわたって，各ケプストラムの値を平均化し，その値を各時点のケプストラムから差し引く方法である．ケプストラム領域での引き算は，スペクトル領域での除算に相当するので，この処理により，平均的なスペクトル特性の逆フィルタが行われることになり，比較的定常的な歪みの影響を除去することができる．簡単ではあるが有効性の高い方法である．

通常の雑音は，音声波形あるいはスペクトルに対して加算的であるので，あらかじめ推定した雑音のスペクトルを音声のスペクトルから差し引くスペクトル減算法がよく用いられる．多くの場合，雑音のレベルが変動するので，雑音スペクトルを過大に減算して，負のスペクトルを生じない注意が必要である．これを防ぐために，逆に音声のモデルの方に雑音スペクトルを加算する方法もある．

音響モデルの領域で適用される基本的な適応化法に，ベイズ適応学習，あるいはそれから導かれる MAP 推定（maximum a posteriori estimation）による方法がある[3]．この方法は，不特定話者モデルのような，信頼性は高いが現在の入力音声には必ずしも合っていないモデルと，現在の入力音声に近いが少量のデータに基づいているために信頼性が低いモデルを，適切に組み合わせて信頼性の高いモデルを作ることができる．

雑音や歪みのある音声への HMM の適応化法に，HMM 合成法（しばしば PMC: parallel model combination と呼ばれる）がある[3]．雑音も HMM でモデル化し，音声の HMM と合成して，雑音が重畳した音声の HMM を作成する方法である．

少量の音声サンプルを用いて適応化をする場合，この中には限られた数の音素しか含まれないので，含まれない音素を推定することが必要になる．この方法としては，スペクトル空間における近傍の音素の変化から補間（平滑化）する方法，それを階層的に行う方法，変化を比較的単純な線形変換（アフィン変換）や少数のバイアスに拘束する方法などが試みられている[3]．

従来の音声認識におけるパラメータ推定は，事後確率あるいは尤度最大化を規準にして行われることが多かった．しかし，音声認識の究極の目的は誤認識の最小化にあるので，この規準を直接的に用いてパラメータを推定する方法，MCE/GPD 法 (minimum classification error/ generalized probabilistic descent algorithm) が，提案されている．相互情報量などに基づいて，不特定話者の音素を識別するのに有効な判別的特徴や距離尺度を構築し，これを用いて認識を行う方法も試みられている．

## 4. むすび

HMMに関連した音声認識の最新技術について解説した．技術の詳細については参考書，論文などを参照していただきたい．音声認識の技術は，近年大きく進歩しているが，任意の話題について，任意の人が，任意の環境で，任意の話し方で話した音声が認識できるという目標からは，まだ程遠い状況にある．この目標に近付くためには，話し言葉の音声データベースを大量に収集して，そのメカニズムを解明し，話し言葉の音声に整合した音響モデルや言語モデルを構築する必要があり，このようなプロジェクトも始まっている[5][6]．

音声認識のプロセスを逆問題として論じた人

はまだいないように思う．逆問題の厳密な定義が何なのかもよく知らないが，冒頭で述べたように，音声認識のプロセスを広い意味での逆問題としてとらえても，間違いではないであろう．逆問題あるいは逆解析の問題としてとらえることによって，現在かかえている多くの問題の解決につながる，新しいアプローチが展開できることを期待したい．

参考文献


[1] 古井貞熙：“音声トランスクリプションのこれまでと今後の展望”，電子情報通信学会論文誌，J83-D-II，11，pp. 2059-2067 (2000)

[2] B.-H. Juang and S. Furui: “Automatic recognition and understanding of spoken language – A first step toward natural human-machine communication”, Proc. IEEE, 88, 8, pp. 1142-1165 (2000)

[3] 古井貞熙：“音声情報処理”，森北出版，東京 (1998)

[4] 中川聖一：“確率モデルによる音声認識”，電子情報通信学会，東京（1988）

[5] 古井貞熙：“話し言葉処理のメカニズム―音声認識から音声理解へ―”，映像情報メディア学会誌，54，6，pp. 765-768 (2000)

[6] 篠崎、齋藤、堀、古井：“話し言葉音声の認識を目指して”，電子情報通信学会技術報告 (2000)