特　集　「時系列データの認識—HMMを越えて」

音声認識においてHMMとトライグラムを越えるもの

Beyond HMM and Trigram for Automatic Speech Recognition

中川　聖一　豊橋技術科学大学情報工学系
Seiichi Nakagawa　Department of Information and Computer Sciences, Toyohashi University of Technology.
nakagawa@slp.ics.tut.ac.jp, http://www.slp.ics.tut.ac.jp/

Keywords: speech recognition, acoustic model, HMM, language model, n-gram.

1. はじめに

音声認識技術は，2万語彙や6万語彙を認識対象とするディクテーションプログラムが商品化され，音声認識の研究は格段の進歩を遂げてきた．大量の音声データベースとテキストデータベースによって音声認識用のモデル（音響モデルと呼ばれる）と言語処理用のモデル（言語モデルと呼ばれる）を構築し，高速・高精度な探索を実現できるようになったことによる．しかし，機械による音声認識能力はまだ人間の能力に遠く及ばない[中川 00a]．

音声認識は，時系列パターン認識の典型的な問題を包含している．つまり，パターンが時間的に非線形に伸縮しており，ある時刻のパターン（スペクトル）そのものもゆれていて，局部的に見ても同一のパターンは存在しない．この点が，遺伝子解析やテキスト解析と大きく異なる．音声認識問題は，音声の生成過程を音響モデルと言語モデルで表現し，これらと入力パターンとの照合問題として定式化される．現在，最も精度の良い認識モデルは，前者を隠れマルコフモデル（Hidden Markov Model: HMM），後者をトライグラムでモデル化するものである．HMMはトライグラム（二重マルコフモデル）を包含し，確率文脈自由文法はHMMを包含するが，他を包含するモデルのほうが適切とは必ずしも言えない点が，パターン認識の難しい点である．

言語モデルで用いられるn-gramは（$n-1$）重マルコフモデルのことで，単純マルコフモデルに変換でき，HMMに包含される．また，線形予測分析（LPC）は音声を定常ガウス過程と仮定しており，p次の分析はp重マルコフ過程であり，これも隠れマルコフモデルに包含される．しかし，包含している抽象度の高いモデルのほうが良いとは必ずしも言えず，対象に特化したモデルを探求する必要がある．例えば，トライグラムや可変グラムの言語モデルを隠れマルコフモデルで表現するためには，膨大な状態数が必要であり，パラメータの推定は非常に難しくなる．

表1は現在の音声認識のアプローチを示している．明らかに，現在の研究の主流は統計的・確率的手法に基づいている．これらの手法に人間の聴覚機構・知覚機構・言語処理（記号操作）機構の知見をいかに組み込むかが研究のトピックになっている．本解説では，HMMおよびトライグラムの欠点とさまざまな改善法・新しいモデルについて述べる．なお，HMMの詳細は[中川 88]を，最近の研究動向は[中川 00a]を参照されたい．

2. 音響モデル

2·1　HMMの原理と改良

今，$Y=y_1, y_2, \cdots, y_T$を音声時系列パターンとしよう．ここで，y_iは第i時間区分（第iフレームという）の音声の特徴を表す特徴ベクトル（通常はスペクトル包絡を表現するパラメータ集合）である．これをベクトル量子化技術によって有限個のシンボルの一つに変換し，Yを離散シンボル系列として扱う場合もあるが，量子化誤差を伴う劣化の問題が生じるので，本解説では連続値のまま扱うモデルについて述べる．このとき，Yを観測して，

表1　音声言語処理のモデル

モデル	レベル			
	音響・信号処理	音声認識	言語処理	言語理解
簡易モデル	自己回帰モデル	隠れマルコフモデル （継続時間，動的特徴）	1文内の文脈自由文法 〃字組確率	テンプレート・フレーム スクリプト
詳細モデル	自己回帰・移動平均モデル 聴覚モデル	確率線形動的システムモデル セグメント統計量	文間にわたる確率文脈依存文法 長距離間の〃字組確率	ユーザの意図 信念維持管理

単語列 $W=w_1, w_2, \cdots, w_n$（音韻列，音節列と考えてもよい）を見い出す問題を考えよう．このとき $P(W|Y)$ を最大にする W を見い出すのが妥当であろう．ここで，

$$P(W|Y)=P(Y|W)\cdot P(W)/P(Y) \quad (1)$$

であるから，$P(Y|W)$，$P(W)$，$P(Y)$ が求まればよい．ここで，$P(Y)$ は最適化しようとしている W とは無関係であるから考慮しなくてよい．

$$P(W)=P(w_1)\prod_{i=2}^{n}P(w_i|w_{i-1}w_{i-2}\cdots w_1) \quad (2)$$

は W の事前生起確率であり，認識対象の言語モデル（文法など）から計算できる．高頻出の単語列は慣用句が多く，発音があいまいになるので特別の音響モデルが必要であろうなど，言語モデルと音響モデルは完全に独立に考察されるものでもない．

$$P(Y|W)=P(y_1y_2\cdots y_T|w_1w_2\cdots w_n) \quad (3)$$

は単語間にわたってはパターンの変動に影響がないとすれば（コンテキスト独立の仮定），

$$\begin{aligned}
&P(y_1y_2\cdots y_T|w_1w_2\cdots w_n)\\
&=\sum_{t_1<t_2\cdots<t_{n-1}}P(y_1y_2\cdots y_{t_1}|w_1)\\
&\quad\times P(y_{t_1+1}y_{t_1+2}\cdots y_{t_2}|w_2)\cdots\\
&\quad\times P(y_{t_{n-1}+1}y_{t_{n-1}+2}\cdots y_{t_n}|w_n)\\
&\cong\max_{t_1<t_2\cdots<t_{n-1}}P(y_1y_2\cdots y_{t_1}|w_1)\\
&\quad\times P(y_{t_1+1}y_{t_1+2}\cdots y_{t_2}|w_2)\cdots\\
&\quad\times P(y_{t_{n-1}+1}y_{t_{n-1}+2}\cdots y_{t_n}|w_n) \quad (4)
\end{aligned}$$

となる（$t_n=T$）．式（4）は，順序を保ったあらゆる時間区間の区切り方に対して右辺の項の総和をとるか，最大値をとるかを表している．結局，任意の $y_iy_{i+1}\cdots y_j$ と任意の w_k に対して $P(y_iy_{i+1}\cdots y_j|w_k)$ が計算できれば求まる．$P(W)$ を求めるためのモデルを言語モデルと呼ぶのに対し，$P(Y|W)$ を求めるためのモデルは音響モデルと呼ばれる．結局，$P(y_iy_{i+1}\cdots y_j|w_k)=P(y_i|w_k)\times P(y_{i+1}|y_iw_k)\cdots P(y_j|y_iy_{i+1}\cdots y_{j-1}w_k)$ であるから，$P(y_n|y_my_{m+1}\cdots y_{n-1}w_k)$ が求まればよいことになる．しかし，$y_m\cdots\cdots y_{n-1}$ は連続値を取るベクトルであるから，モデルのパラメータが多くなり容易に計算できない．そこで，媒介変数 $x_i, x_{i+1}\cdots x_j$（状態変数）を用いて，その系列を x で表現すれば以下のように近似する（w_k を省略した）．

$$\begin{aligned}
&P(y_1\cdots y_t)\\
&=\sum_x P(y_1y_2\cdots y_t, x_1x_2\cdots x_t)\\
&=\sum_x P(y_1y_2\cdots y_t|x_1x_2\cdots x_t)P(x_1x_2\cdots x_t)\\
&=\sum_x\prod_i P(y_i|y_1y_2\cdots y_{i-2}y_{i-1}\ x_1x_2\cdots x_{i-1}x_i)\\
&\quad\times P(x_i|x_1x_2\cdots x_{i-1}) \quad (5)\\
&\cong\sum_x\prod_i P(y_i|y_{i-3}y_{i-2}y_{i-1},\ x_{i-1}x_i)P(x_i|x_{i-1})\\
&=\sum_x\prod_i\frac{P(y_{i-3}y_{i-2}y_{i-1}y_i|x_{i-1}x_i)}{P(y_{i-3}y_{i-2}y_{i-1}|x_{i-1}x_i)}P(x_i|x_{i-1}) \quad (6)\\
&\cong\sum_x\prod_i P(y_i|x_{i-1}x_i)P(x_i|x_{i-1}) \quad (7)
\end{aligned}$$

式（5）から式（6）への変形には x の単純マルコフ性を仮定し，式（7）への変形には y 間の独立性を仮定している．式（7）が隠れマルコフモデルと呼ばれるもので，第一項が出力確率，第二項が状態遷移確率である．通常，第一項は，多次元正規分布の混合でモデル化する．第二項は，同じ状態に遷移が続くと指数関数的に確率は減少する．式の変形から明らかなように隠れマルコフモデルは原式から見れば第1次程度の近似であり良くないことがわかる．それにもかかわらず，パラメータ数が比較的少なく学習が容易（EMアルゴリズムによる最尤推定に基づくBaum-Welchアルゴリズム．Forward-backwardアルゴリズムとも呼ばれている）であることから，広く一般に用いられてきた．第一項の近似の欠点を補うために「動的特徴パラメータ」が付加され，第二項の近似の欠点を補うために「継続時間制御モデル」が導入されている．フレーム間の動的変化を考慮するための手法としては，HMMの状態数を増やす方法や，特徴パラメータの時間軸方向の1次と2次の回帰係数「Δ, $\Delta\Delta$ パラメータ」を用いる方法，線形や非線形の予測器を使用する方法，多項式回帰関数を使用する方法，セグメントの統計量を用いる方法，条件付きHMMを使用する方法，人間の聴覚特性の順向マスキングの時間–周波数特性を模擬した動的ケプストラム係数の使用といったような数々の方法が従来より研究されてきている [中川 00a].

我々の最近の研究では，より近似の少ない条件付き出力確率 $P(y_t|y_{t-3}y_{t-2}y_{t-1}x_{t-1}x_tw_k)$ よりも $P(y_{t-3}y_{t-2}y_{t-1}y_t|x_{t-1}x_tw_k)$ を用いるのが従来のどの方法よりもよいことを見い出している．つまり，4フレームのセグメントの統計量を用いるのが有効である [中川 00b].

ほかにセグメントを利用した方法としては，ニューラルネットワークを用いて連続する数フレームを一括して入力し，その出力値をHMMの入力とする方法も数多く試みられてきた．また，最近になって条件付きHMMの改良や，セグメントをHMMの入力とする方法が数多く試みられるようになってきている [中川 00a].

HMMの状態 i に割り付けられる観測ベクトル $y_1y_2\cdots y_T$ のフレーム間の相関を混合多項式回帰モデルでモデル化するDengらの方法は次式で与えられる [Deng 97]（一般的に，トレンドモデルと呼ばれている）．

$$y_t = \sum_{r=0}^{R} B_{i,m}(r)(t-\tau_i)^r + N_t(0, \Sigma_i) \qquad (8)$$

ここでは，iは状態，mは混合分布のインデックス，τ_iは状態に割り付けられた最初の時間を表す．通常のHMMは$R=0$に相当し，状態iでのy_tの分布を平均ベクトル$B_{i,m}(0)$，共分散行列Σ_iのガウス分布$N_t(B_{i,m}(0), \Sigma_i)$とみなすことになる．$R=0$よりも$R=1$のほうが，$R=1$より$R=2$のほうが良い結果が得られている．このモデルにおいても1次回帰係数の併用は効果がある．ただし，コンテキストに独立な音韻モデルでは効果があるが，コンテキスト依存の音韻モデル（トライフォン）などではパラメータが多くなり過ぎ，その効果は小さい．

Holmesらは，式（8）はセグメント内のパラメータの動的変化のモデルであり（セグメント内変動），このほかに，このモデル自身$\{B_{i,m}\}$が変動する（セグメント外変動）というモデルに拡張している（$R=1$，$m=1$の場合について[Holmes 99]）．この方法によって，Δケプストラムを併用したバイフォンモデルにおいても認識性能が若干向上した．さらに，最近，これを混合モデル化したものが試みられている[Xiang 01]．

継続時間制御のためには，各状態ごとの継続時間分布をガウス分布，ガンマ分布，ポアソン分布，多項分布（離散分布）でモデル化する方法が試みられてきた．より厳密には，各状態ごとと各モデルごとの2種類の分布，およびモデル間の分布の相関なども用いる必要がある．なお，状態遷移確率を状態の滞在時間長によって可変にするモデルが提案され状態継続時間分布を用いるよりも良い結果が報告されている．また，出力確率を状態継続時間長に依存する方法も提案されている[中川 00a]．

2·2 HMMを越えるモデル

Ostendorfらのセグメントモデルによる音声認識は次のように定式化される[Ostendonf 96]．

$$P(y_1 y_2 \cdots y_T | w_1 w_2 \cdots w_N)$$
$$= \sum_{l_1^N} P(y_1^T, l_1^N | w_1^N)$$
$$= \sum_{l_1^N} P(y_1^T | l_1^N, w_1^N) P(l_1^N | w_1^N) \qquad (9)$$

$$P(y_1^T | l_1^N, w_1^N) = \prod_{i=1}^{N} P(y_{t(i-1)+1}^{t(i)} | l_i, w_i) \qquad (10)$$

$$P(l_1^N | w_1^N) = \prod_{i=1}^{N} P(l_i | w_i, l_{i-1}, w_{i-1}) \qquad (11)$$

ここで，l_iは単語（音韻）w_iの時間長である．l_iをさらにいくつかの区間に分けてモデル化すると，この区間は状態に対応し，単純化した場合はHMMと同様のモデルになる．式（11）はセグメンテーションの確率を表しており，認識に重要であると言われている．ただし，式（10）の求め方（通常は固定次元に変換する）にも依存するが，フレーム間の相関は考慮したモデルになっている．しかし，相関まで考慮するとパラメータ数が多くなり過ぎ，良くならないこともある．このようなセグメントモデルとHMMの併用が現在のところ最も高精度な音響モデルであると考えられる[高橋 01]．

HMMやセグメントモデルを拡張した次の確率線形動的システム（状態空間動的モデル）の適用も試みられている[Digalakis 93]．

$$x_{t+1} = F_t x_t + G_t w_t \qquad \text{（状態方程式）}$$
$$y_t = H_t x_t + v_t \qquad \text{（観測方的式）} \qquad (12)$$

ここで，x_tは非観測の状態ベクトル（連続値），y_tは観測ベクトル，w_t, v_tはx_t, y_tと無相関な平均値0のガウス確率変数ベクトル，F_tは状態遷移行列，G_tは駆動行列，H_Tは観測行列である．一般に$\{x_t\}$はガウス・マルコフ過程で$\{y_t\}$はガウス過程であるがマルコフ過程でない．通常のHMMやARモデルは上記の特殊例（時不変システム）としてモデル化される．

通常のカルマンフィルタは$\{F_t\}$, $\{G_t\}$, $\{H_t\}$, x_oなどは事前に既知として，$\{y_t\}$からx_tを推定するものである．音声認識の場合は，$\{y_t\}$から$\{F_t\}$, $\{G_t\}$, $\{H_t\}$などを推定する必要がある．これらのパラメータはセグメントに対応するモデルごとに時不変と考えられる区間（HMMの状態に相当）ごとに求める必要がある．本モデルは，統計的セグメントモデルのセグメント内の相関を考慮した方法の近似モデルになっており，Δケプストラムよりも効果は小さいが，その併用効果はある．Afifyらは，Digalakisらのモデルを複数の動的変化を扱えるように発展させている（混合モデル）[Afify 96]．Dengらはx_tのターゲットを声道共振周波数（第1～第3フォルマント周波数）に対応させ，H_tをニューラルネットにより実現する非線形写像関数で定義する手法を提案している[Deng 00]．なお，HMM，セグメントモデル，動的システムの関係は文献[Ostendonf 96]で述べられている．

HMMを包含するベイジアンネットワークも音声認識に用いられている[Zwing 98]．DevirenとDaoudiの提案した動的ベイジアンネットワークは次のように定式化される[Deviren 01]．$x_1^T = \{x(1), x(2), \cdots, x(T)\}$, $x(t) = \{x_h(t), x_o(t)\}$とする．ここで，$x_h(t)$と$x_o(t)$は時刻$t$における隠れ変数と観測ベクトルとする．観測ベクトルには，時間的に動的なパラメータである，Δ, $\Delta\Delta$パラメータを含むとする．

$$P(x_h(t) | x_1^{t-1})$$
$$= P(x_h(t) | x_h(t-k), \cdots, x_h(t-1))$$
$$P(x_o(t) | x_1^T \setminus \{x_o(t)\})$$
$$= P(x_o(t) | x_h(t-\tau_p), \cdots, x_h(t-\tau_f))$$

$A \setminus B$は集合Aから集合Bを除くことを意味する

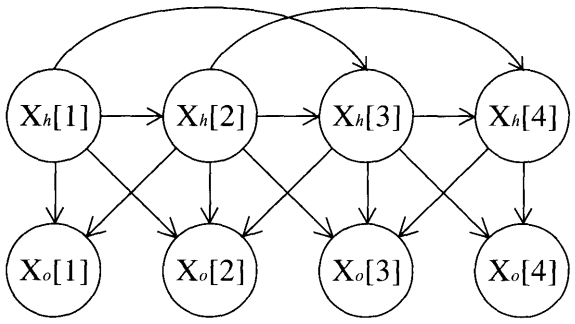

図1 動的ベイジアンネットワークの例

通常のHMMは $(k, \tau_p, \tau_f)=(1, 0, 0)$ に対応する．図1は $(k, \tau_p, \tau_f)=(2, 1, 1)$, $T=4$ の場合の動的ベイジアンネットワークの構造を示している．予備的な数字音声の認識実験で，通常のHMMで92.9％，(1,0,1) で96.4％，(2,1,1) で98.4％の認識率を得ている．HMMの状態数を多くすれば，同様の構造が実現できると思われるが，ベイジアンネットワークの構造のほうが状態固定後の状態間の相関や将来のサンプルを用いているなどより一般化されている．

このほか，HMMを包含するモデルとして，連続確率文脈自由文法 [中川 92]，マルコフランダム場 [Zhao 91]，リカレントニューラルネット [Robinson 94] などが試みられているが，音素認識や単語認識においては，一部HMMよりも優位な報告もあるが，連続音声認識ではHMMを優位に上回るには至っていない．

2·3 音響モデルの単位

音声パターンは前後のコンテキストによって影響を受ける（調音結合）．そのため，この影響を除いたり，正規化する研究は数多く行われてきたが成功しなかった．そこで，さまざまなコンテキストの影響を受けたパターンをコンテキストを考慮せずにモデル化するのは限界があることから，コンテキスト別にモデル化するアプローチが主流となっている．前後の音韻を考慮しない音韻単位よりも前後の音韻別の音韻単位を用いたほうが認識性能は良いことがわかろう．テンプレートマッチングでは，母音–子音–母音のコンテキスト別に標準パターンを用いる方法が試みられていたが，HMMではコンテキスト別に音韻パターンを用いるのが主流である．コンテキスト別に音韻パターンを用いる方法は最初BBN社で検討され，CMUのK-F. Leeの研究によって有効性が広く知られるようになりトライフォンモデルとして定着した [Lee 89]．K-F. Leeは，単語間にわたるコンテキストや前置詞などの発声時間の短い機能語に対しても特別な音韻モデルを用いるなど，音響モデルを精密化した．さらに，コンテキストの影響の大きい場合には，前後2音韻のコンテキストまで考慮したquinphoneモデルも導入されている [Young 97]．

3. 言語モデル

音声認識における言語情報の役割は音声言語を理解し，正しく動作することを第一主義とすることは言うまでもないが，実際上は探索空間を小さくし，いわゆる非文が認識結果とならないように排除し，認識精度を向上させることである．すなわち，ターゲット言語を包含するできるだけ小さい文集合を生成する言語モデルが望ましい．現在の音声認識技術では，音声認識結果のあいまいさは避けられず，音響レベルだけでは候補となり得る可能な認識結果は無限と考えてよい．探索空間を小さくするということは，良い言語モデルを構築することであり，良い言語モデルとは認識対象の文集合のエントロピー（パープレキシティ*1）を小さくするモデルである．このことから，認識対象の文集合に存在する統計的性質を利用するのが自然である．これが言語の確率モデルである．音声ワープロのように書かれた文を音声で入力する場合は，文法を中心とした文字言語の言語モデルで十分と思われるが，話し言葉である音声言語では，言い淀み，言い直し，助詞落ち，倒置，間投詞の挿入など，文字言語では非文扱いとなる文が頻繁に生じるため，言語モデルの構築は極めて難しい（表2参照）[中川 00a]．例えば，「富士山観光案内」という限定されたタスクにおいてさえ，これに関して発声される文集合を受理する文脈自由文法を作成することは非常に難しく，カバーレイジは約40～80％程度であり，パープレキシティも100前後と大きくなる．この場合でも，数百文の学習データから単語単位の（クラス）バイグラムを用いるとカバーレイジも大きくパープレキシティも小さくなる [中川 98]．

表2 言語モデルの比較

構文解析用の評価尺度：解析木の良否
音声認識用の評価尺度：エントロピーの大小

アプローチ	対象	人間による規則生成	機械による規則（モデル）生成
非統計的	構文解析	△	×
	音声認識	△	×
統計的	構文解析	◎	○
	音声認識	○	◎

形式的に言えば，言語モデルは，与えられたターゲット言語において，単語列 $w_1, w_2, \cdots, w_n$ に隣接する w_{n+1} の条件付き生成確率 $P(w_{n+1}|w_1, w_2, \cdots, w_n)$ を求める確率モデルのことである．これは，$w_1, w_2, \cdots, w_n$ の同時生起確率 $P(w_1, w_2, \cdots, w_n)$ などを求めるモデルと同質であり，一般に前者は $(n+1)$ 重マルコフモデル，後者

*1 情報理論的な意味での後続可能単語候補数．2のエントロピー乗で定義される．

はn-gramと呼ばれている．3重マルコフモデル（4-gram）などの算出には多量のデータと記憶量が必要であり，実質上不可能である．そこで，2重マルコフモデル（tri-gram）や文法などによって近似することになる．ある程度タスクやドメインが限定されている場合，3gramによるパープレキシティは50～200程度となる．

ShannonやCoverらのテキストの読取りによる後続アルファベットの予測に関する被験者実験から，英語の単語単位のパープレキシティは100前後（1文字当たり1.3ビット）と予想されているので[中川 88]，3gramによる言語モデルはかなり精度の高いことがわかる．3gramを改良して定型表現や4gramを導入してもパープレキシティの減少は約10％であり，この程度の改善は単語認識率の向上にあまり貢献していないと考えられる[中川 00a]．

単語列の局所的な出現頻度だけでなく，もっと長いスパンを考慮する言語モデルとしてトリガーモデルがある[Rosenfeld 96]．例えば「私は彼と一緒に学校へ通う」と「私は学校へ彼と一緒に通う」の場合を考えてみよう．「学校」と「通う」は関連する共起しやすい組であるが，前者では3gramで考えることができるが後者では7gram必要になる．このように，ある単語が出現するとこれがトリガーとなって，継続する3gramを修飾する方法であり，10％程度のパープレキシティの削減をもたらす．

最近では，言語モデルをより精密化するために，頻出する単語系列の一部に4gramを使用したり，フレーズや定型表現を一単位として扱ったり，可変長n-gramを併用するなどのより長いコンテキストを考慮する試みが増えている．

可変n-gramの素直な記述法としてPSM（prediction suffix tree）による言語モデルが提案され，3gramよりも10％前後パープレキシティの減少が報告されている[Pereira 99]．

なお，4～5gram，単語クラス単位の言語モデル，直前の発話文による言語モデルの適応（キャッシュモデル），種々の文タイプ別の言語モデルの混合などを統合することにより30～40％前後のパープレキシティの減少が報告されている[Goodman 01, Martin 97]．

一方，句構造文法のような規則に基づく方法では，話し言葉のような多様な言語現象をカバーし非文を排除することは極めて難しい．当然，規則に確率を付与する確率文法が考えられるが，元となる規則を作成する困難さは変わらない．そこで，規則自体を学習する試みがなされている．

自然言語の比較的良いモデルと言われている文脈自由文法（CFG）に対しては，大量データからの確率文脈自由文法の学習法はInside-Outsideアルゴリズムとして知られている（確率文脈自由文法よりも記述能力の低い確率正規文法は，ほぼHMMと等価と考えてよい）．しかし，この学習は繰返し計算量が多いこと，局所的最適解に収束することなどにより難しい．そこで，括弧付き（構文木付き）コーパスを制約として用いる方法などが提案されている[Pereira 92]．CFGの規則の確率化よりも，導出途中の部分木の出現確率を用いるほうが解析精度は向上するが，音声認識への適用化は難しく，3gramよりもパープレキシティはなかなか良くならない．

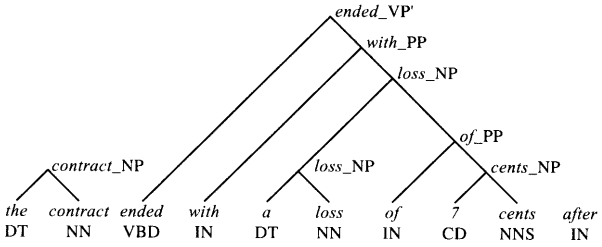

図2　構造化言語モデルで使用される部分構文解析木の例

最近，ChelbaとJelinekらは次の構造化言語モデルを提案した[Chelba 00]．図2は“the contract ended with a loss of 7 cents after”の構文解析結果例である．“after”を予測するには3gramによる直前の2単語ではなく，直前の句の主辞（head word）である“ended”が必要である．このため，構文解析結果を用いる次のモデルが提案された．

$$
\begin{aligned}
&P\left(W_1^n, T_1^n\right)\\
&=\prod_{i=1}^{n} P\left(w_i \mid h_{-2}, h_{-2}.tag, h_{-1}, h_{-1}.tag\right)\\
&\times P\left(w_i.tag \mid h_{-2}.tag, h_{-1}.tag, w_i\right)\\
&\times P\left(T_1^i \mid w_i, w_i.tag, T_1^{i-1}\right)
\end{aligned}
$$

ここで，$W_1^n = w_1 w_2 \cdots w_n$は入力文，$T_1^n$はそれに対応する構文解析木である．$h_{-m}$は$m$個前の主辞，$h_{-m}.tag$はそのタグ（品詞）を示す．このモデルは，単独でもトライグラムよりやや良く（パープレキシティで0～5％減少），トライグラムとの併用で，パープレキシティを約10％減らすことができている．

4. むすび

本解説では，音声認識のためのHMMを中心とした音響モデル，3gramを中心とした言語モデルに関して，最近の研究動向を述べた．静かな部屋で比較的丁寧に発声すれば高精度に音声認識ができるようになってきた．しかし，だれもが動きながら自由に発声した音声を認識するためには，まだまだ基礎研究が必要である．音韻認識率の2～3％の改善とパープレキシティの20～30％の減少は単語認識率には，ほぼ同等の改善効果をもたらす[中川 00]．これから，音声分析や特徴パラメータ，言語モデルの研究よりも音響モデル（動的特徴パラメー

タの扱い）と認識アルゴリズムの研究が重要だと考えられる．これには，人間の知覚過程の解明が手掛かりを与えてくれるであろう．なぜなら，現在主に用いられる特徴パラメータから音声を再生すれば，我々はほぼ正しく認識・理解できるし，ある程度タスク（ドメイン）が限られていれば，人間の単語予知能力とn-gramの適応化による単語予測能力にはそれほど大きな差がないからである．

◇ 参 考 文 献 ◇

[Afify 96] M. Afify, Y. Gong and J-P. Haton: Estimation of mixtures of stochastic dynamic rejectories: Application to continuous speech recognition, Computer Speech and Language (1996)

[Chelba 00] C. Chelba and F. Jelinek : Structure language modeling, Computer Speech and Language, Vol. 14, No. 4, pp. 283-332 (2000)

[Deng 97] L. Deng and M. Aksmanovic: Speaker-independent phonetic classification using hidden Markov models with mixtures of trend functions, IEEE Trans. Speech and Audio Process, Vol. 5, No. 4, pp. 319-324 (1997)

[Deng 00] L. Deng and J. Ma: Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics, J. Acoust. Soc. Am., Vol. 104, No. 6, pp. 3036-3048 (2000)

[Deviren 01] M. Deviren and K. Daoudi: Structual learning of dynamic Bayesian networks in speech recognition, Proc. EuroSpeech, pp. 1669-1672 (2001)

[Digalakis 93] V. Digalakis, J. R. Rohlicek and M. Ostendorf: ML estimation of a stochastic linear system with the EM algolithm and its application to speech recognition, IEEE Trans. Speech Audio Process, Vol. 1, pp. 431-442 (1993)

[Goodman 01] J. T. Goodman: A bit of progress in language modeling, Computer Speech and Language, Vol. 15, pp. 403-434 (2001)

[Holmes 99] W. J. Holmes and M. J. Russel: Probabilistic-trajectory segmental HMMs, Computer Speech and Language, Vol. 13, pp. 3-37 (1999)

[Lee 89] K. F. Lee: Automatic Speech Recognition. The development of the SPHINX system, Kluwer Academic Publishers (1989)

[Martin 97] S. Martin, J. Liermann and H. Ney: Adaptive topicdependent language modeling using word-based varigrams, Proc. Eurospeech, pp. 1447-1450 (1997)

[中川 88] 中川聖一: 確率モデルによる音声認識, 電子情報通信学会 (1988)

[中川 92] 中川聖一: 連続出力分布型確率文脈自由文法の提案, 日本音響学会春季大会, 3-1-2 (1992)

[中川 98] 中川聖一, 大谷耕嗣: bigramの使用による話し言葉用確率文脈自由文法の自動学習, 情報処理学会論文誌, Vol. 39, No. 3, pp. 575-584 (1998)

[中川 00a] 中川聖一: 音声認識研究の動向, 電子情報通信学会論文誌, Vol. 83-D, No. 2, pp. 433-457 (2000)

[中川 00b] 中川聖一, 花井建豪, 山本一公, 峯松信明: HMMに基づく音声認識のための音節モデルとtriphoneモデルの比較, 電子情報通信学会論文誌, Vol. 83-D, No. 6, pp. 1412-1421 (2000)

[Ostendonf 96] M. Ostendonf, V. V. Digalakis and O. A. Kimball: From HMMs to segment models: A unified view of stochastic modeling for speech recognition, IEEE Trans. Speech and Audio Process, Vol. 4, No. 5, pp. 360-378 (1996)

[Pereira 92] F. Pereira and Y. Schabes: Inside-Outside reestimation from partially bracketed corpora, Proc. ACL, pp. 31-37 (1992)

[Pereira 99] F. Pereira, Y. Singer and N. Tshby: Beyond word N-grams, Proc. Natural Language Processing Using Very Large Corpora, pp. 121-136 (1999)

[Rosenfeld 96] R. Rosenfeld: A maximum entropy approach to adaptive statistical language modeling, Computer Speech and Language, Vol. 10, pp. 187-228 (1996)

[Robinson 94] A. J. Robinson: An application of recurrent nets to phone probability estimation, IEEE Trans. Neural Networks, Vol. 5, No. 2, pp. 298-304 (1994)

[高橋 01] 高橋伸寿, 中川聖一: 音節セグメントの統計量を用いた音節認識, 日本音響学会講演論文集, 1-1-7 (2001.10)

[Xiang 01] B. Xiang and T. Berger: Multiple mixture segmental HMM and its application, Proc. ICALLP, pp. 509-512 (2001)

[Young 97] S. J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J. L. Gauvain, D. J. Kershaw, L. Lamel and D. A. Leeuwen: Multilingual large vocabulary speech recognition in the European SQALE project, Computer Speech and Language, Vol. 11, pp. 73-89 (1997)

[Zhao 91] Y. Zhao, L. E. Atlas, and X. Zhuang: Application of the Gibbs distribution to hidden Markov modeling in speaker independent isolated word recognition, IEEE Trans. Signal Process, Vol. 39, No. 6, pp. 1291-1298 (1991)

[Zwing 98] Z. Zwing and S. Russel: Probabilistic modeling with Bayesian networks for automatic speech recognition, Proc. ICSLP, pp. 3011-3014 (1998)

2001年11月8日 受理

著 者 紹 介

中川 聖一（正会員）

1976年京都大学大学院博士課程修了．同年京都大学助手．1980年豊橋技術科学大学情報工学系講師．1983年助教授．1990年教授．音声処理，自然言語処理，人工知能の研究に従事．1977年度電子通信学会論文賞，1988年度IETE最優秀論文賞，2001年度電子情報通信学会論文賞．著書：「音声・聴覚と神経回路網モデル」（共著），「確率モデルによる音声認識」，「情報理論の基礎と応用」，「音声」（共著），「パターン情報処理」など．