■2 群（画像・音・言語）- 7 編（音声認識と合成）

2 章 音声認識

（執筆者：河原達也）[2009 年 8 月 受領]

■概要■

音声認識の研究は今から 50 年以上前にさかのぼるが，1990 年代以降，HMM や n-gram に代表される統計的モデルに基づく方法論の確立と，DARPA などの国家プロジェクトによる大規模なコーパスの構築，そして計算機性能の飛躍的増大が相乗するかたちで大きく発展した．その結果，数万の語彙を対象とした連続音声認識も実現され，ディクテーションや自動電話応答，音声翻訳，会議録作成支援などのアプリケーションに展開されるに至っている．

音声認識は以下のように，入力音声の特徴量 X に対して事後確率 $p(W/X)$ が最大となる単語列 W を見つける問題として定式化され，図 2・1 のような構成となる．

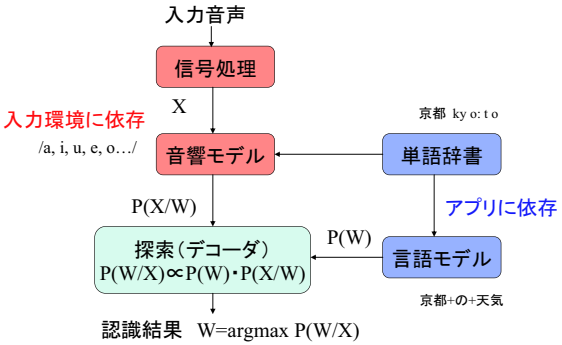

図 2・1 音声認識システムの概要

図 2・1 で，入力音声から周波数分析により特徴量 X を求めるのが信号処理・音声分析である．音響モデルは，音素などの単位で標準的な特徴量パターンを保持しておき，入力と照合して尤度 $p(X/W)$ を与えるもので，一般的に HMM が用いられる．単語辞書は，認識対象の語彙（＝単語の集合）とその発音（＝音素の系列）を規定するもので，この単語辞書を照合しながら音響モデルによる音素単位の認識が連続的に実行される．言語モデルは単語の連鎖に関する制約や尤度 $p(W)$ を規定するもので，文法規則あるいは n-gram モデルなどに基づいて構成される．これらのモデルを統合して最尤の単語列仮説を探索するプログラムがデコーダ，あるいは認識エンジンと呼ばれる．以下の節では，これらの各モジュールについて解説する．

信号処理と音響モデルは入力環境に依存し，単語辞書と言語モデルはアプリケーションのタスクドメインに依存するので，これらに応じて用意・設定する必要がある．特に，雑音が多い環境では特別の対応が必要となるし，話者や環境に音響モデルを適応させるのも効果的であるので，これらについても独立した節で解説する．

【本章の構成】

本章は以下，音声認識の概要（2-1 節），音響・音素モデル（2-1 節），言語モデル（2-2 節），大語彙連続音声認識アルゴリズム（2-3 節），話者・環境適応（2-4 節），雑音に頑健な認識（2-5 節），話し言葉音声の認識・処理（2-6 節），話者認識・インデキシング（2-7 節）から構成される．

■参考文献

1) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，“音声認識システム.,” オーム社，2001.
2) 河原達也, 李晃伸, “連続音声認識ソフトウエア Julius.,” 人工知能学会誌, vol.20, no.1, pp.41-49, 2005.

■2 群 - 7 編 - 2 章

2-1 音響・音素モデル

（執筆者：篠田浩一）[2009 年 7 月 受領]

2-1-1 隠れマルコフモデル

音声認識においては隠れマルコフモデル（hidden Markov model; HMM）が主流である[3)]．時系列データのモデル化には，状態空間モデルと呼ばれるダイナミックベイジアンネットワークがよく用いられる（図 2・2）．その中でも隠れ（潜在）変数が離散変数（状態と呼ぶ）であるモデルを隠れマルコフモデルと呼ぶ．HMM は，ラベル Y がデータ X を生成する確率 $P(X|Y)$ を出力する生成モデルである．それ以前に主流であった DP マッチング（Dijkstra のアルゴリズム）に対し，パスごとに異なるコストを導入し，ラベルの出現を決定的ではなく確率的にしたものととらえることもできる．また，それと等価な重み付き有限状態トランスデューサーに変換できる．DP マッチングと同様に分枝限定法の原理に基づく認識アルゴリズム（Forward アルゴリズム）を持ち，列の長さに対し多項式時間の計算コストで認識処理が可能である．また，Forward アルゴリズムを近似した Viterbi アルゴリズムもしばしば用いられる．

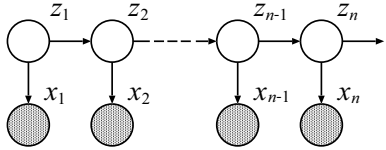

図 2・2　状態空間モデル

$x_1, \ldots, x_n$ は観測変数，$z_1, \ldots, z_n$ は隠れ変数であり，添え字は時刻を表す．隠れ変数が離散変数である場合が隠れマルコフモデル．

HMM のパラメータは，状態出現確率，状態遷移確率，初期状態確率の 3 種類である．状態出現確率はその状態から与えられた観測値を出現する確率であり，その分布としては，多項分布などの離散分布や，ガウス分布，混合ガウス分布などの連続密度分布が用いられる．遷移確率は，ある状態から別の状態へ遷移する確率，初期状態確率は初期時刻に各々の状態に存在する確率を表す．遷移確率，初期状態確率の分布としては，一般に多項分布が用いられる．

HMM の学習では，あらかじめラベルが付与された学習データを用いて，その生成確率 $P(X|Y)$ を最大にする尤度最大化の基準のもと，パラメータ推定を行う．Expectation-Maximization（E-M）アルゴリズムの一種である，Baum-Welch アルゴリズムが一般に用いられる．このアルゴリズムでは，出力と隠れ変数の同時確率最大化の基準のもと，隠れ変数の期待値推定と，尤度最大化基準によるパラメータ推定の繰り返しにより，局所最適解を求める．

2-1-2 HMM による音声認識

現在，HMM は音声認識において広く用いられている．その理由としては，音声に関する知識をその構造に比較的容易に反映させることができること，様々な要因により生ずる音声の変動に対し比較的頑健であること，計算効率が高く，現在の計算機の性能で実時間（発声

長と同程度の時間）以下で認識処理が可能であること，などが挙げられる．HMM による音声認識においては，あらかじめ用意された複数の認識結果候補（単語など）に対して，別々のモデルを用意しておき，認識時には，それらの中から入力発声を出力する確率の最も高いモデルに対応する候補を認識結果とする．

認識単位としては，単語，サブワード（音節，音素など）が用いられる．単語を認識単位とした HMM は単語 HMM，音素を単位とした HMM は音素 HMM と呼ばれる．連続 10 数字の認識など，小規模なタスクでは単語が認識単位として用いられることが多い．大語彙連続音声認識においては，個々の単語の HMM を構成することが困難なため，専らサブワード単位が用いられ，特に音素単位が用いられることが多い．この場合，単語や文の HMM はサブワードを連結して作成される．認識単位の構造としては，音声の性質（単調性，連続性）を考慮し，過去の時刻に通過した状態には遷移をしない，left-to-right 型が用いられる．その場合，スキップあり／なしの両方が考えられる．なお，音声認識のための音声分析のフレーム間隔は，通常 10 ms ほどであり，その場合，音素 HMM の状態は 3 程度とすると認識性能が良いことが経験的に知られている（図 2・3）．

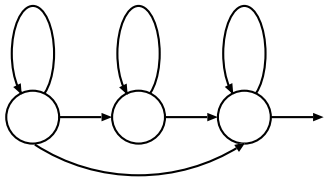

図 2・3　音素 HMM の例として，left-to-right 型スキップありの 3 状態 HMM を示す．この図は状態遷移図と呼ばれ，各○印が状態を示し，その間の矢印が状態遷移を示す．

HMM の状態出力分布としては，混合ガウス分布が主流である．離散分布を用いる場合と比べ計算量は大きいが，音声に含まれる言語情報以外の情報，例えば，話者の違い，周囲雑音の違い，に起因する音響特徴の変動に対し頑健であるという利点がある．

音素などサブワード単位の HMM を学習する場合には，連結学習が用いられる．そこでは，Baum-Welch アルゴリズムの繰り返しステップにおいて以下を行う．まず，各々の発声サンプルに対し，その書き起こしに対応する HMM をサブワード HMM を連結することで作成し，各発声における隠れ変数の期待値の推定を行う．そして，各々のサブワード HMM について，すべてのサンプルにおける期待値を集め，それを用いて尤度最大化を行い，パラメータを更新する．

2-1-3 大語彙連続音声認識における音響モデル

大語彙連続音声認識における音声モデルとしては，HMM などの音響モデルと n-gram などの言語モデルが必要である．ここでは，音響モデルについて述べる[2)]．

大語彙連続音声認識においては，語彙数が大きいため，小語彙の場合に比べ，より精緻な音響モデルが求められる．音素単位（monophone）HMM では，調音結合，すなわち，音素間の渡りにおける発声変形，により生ずる異音（allophone）への対処が難しい．そこで，前後

の音素を考慮した文脈依存音素が認識単位として用いられることが多い．前または後ろの音素どちらか一つを考慮した diphone，前後両方の音素を考慮した，トライフォン（triphone）などがある．一般にはトライフォンがよく用いられる．例えば，先行音素が “b” であり，後続音素が “k” である音素 “a” に対応するトライフォンを “b-a+k” と書く．

トライフォンの種類数は，単純計算では，音素種類数の 3 乗となる．例えば音素が 40 種類ある場合，6 万 4 千種類となる．音素の接続に対する制約などを考慮しても 1 万程度の種類となる．このままではモデルパラメータ数が多くなりすぎ，頑健なパラメータ推定が難しい．また，同一の音素の異音は，ある程度までは同一の音響的性質を共有していることが期待できる．そこで，パラメータ推定の前に，音響的類似性を基準として，パラメータのクラスタリングを行い，自由パラメータ数を減らす処理がしばしば行われる．

クラスタリング法の主流は，音素文脈決定木を用いた状態分割である．この手法は，決定木を用いた尤度最大化手法の一つである．同一の音素を中心音素としてもつすべてのトライフォンにおいて，同じ位置（前から何番目など）にある状態の集合に対しクラスタリングを行う．手順を以下に示す．

1. あらかじめ，弁別素性などに基づく質問の集合 Q を用意しておく．質問としては，右側の音素が閉鎖音である，左側の音素が子音である，などがある．

2. すべての状態を集めた集合を作成し，S とする

3. S を Q の各々の質問に対する答え（Yes, No）で二つにいったん分割してみて，分割後の尤度が最大になる質問 q を求める．

4. 質問 q で S を分割し，その結果の二つの集合をそれぞれ S_1, S_2 とする．

5. S_1, S_2 のそれぞれに対し，3, 4 の処理を繰り返す．

分割後の尤度から分割前の尤度を引いた尤度差は常に非負なので，状態数の上限や尤度差の下限などの閾値を設定し，その閾値に達した時点で分割を停止する．同一のクラスタに属する状態は同じパラメータを共有する．

この音素決定木を用いた状態分割は，トップダウンクラスタリング手法の一つである．文脈依存音素を認識単位として用いる場合，音素の組合せの種類が多いため，学習データ中に全く出現しないトライフォンやほんの少ししか出現しないトライフォンがある．この方法では，弁別素性などの音声に関する事前知識を利用することにより，それらのトライフォンに対するモデルパラメータの推定が，比較的頑健に行われるという利点がある．

2-1-4 識別学習

前項までに述べたように，隠れマルコフモデルは，ラベル Y がデータ X を生成する確率 $P(X|Y)$ を出力する生成モデルである．一方，確率的パターン認識では，データ X を与えた時のラベル Y の条件付き確率 $P(Y|X)$ を最大にしたい．そこで，音声認識では，以下のベイズの定理に基づいて逆問題を解き，事後確率 $P(X|Y)$ を最大化するラベルを求める．

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)} \quad (2\cdot1)$$

ここで，$P(Y)$ はラベル Y に対する言語モデルの出力確率である（本節では一定であると仮定する）．前述の Baum-Welch アルゴリズムでは，$P(X|Y)$ を最大化することを目的とし，右辺の分母は無視していた．しかし，事後確率 $P(X|Y)$ の最大化では，右辺の分子を大きくすると同時に，分母を小さくすることも有効である．この事後確率最大化は，正解を 0，誤りを 1 とした 0-1 損失基準を用いたときの期待損失最小化となり，その意味で，分類誤り個数の最小化を行うための基準とみなすことができる．なお，分類誤りの個数を直接最小化する基準に基づく学習手法も存在するが、事後確率最大化に基づく手法と実用上ほとんど性能差がない．ここでは両者をまとめて識別学習と呼ぶ[3]．

識別学習手法として代表的なものとして，相互情報量最大化（Maximum Mutual Information; MMI）学習，誤識別最小化（Minimum Classification Error; MCE）学習，単語誤り最小化（Minimum Word Error）学習，音素誤り最小化（Minimum Phone Error）学習の四つがある．これらの間の違いは，主に識別の対象となる単位，すなわち，認識率を測る単位による．MCE 学習は文（発声）の識別が対象であり，文単位の認識率（文認識率）の向上を目的とする．同様に，MWE は単語認識率，MPE は音素認識率の向上を目的とする．ちなみに，MMI 学習は，学習音声セットの（他の音声セットに対する）識別を目的とする．違う単位で測った認識性能の間には強い相関があるので，これらの手法間の性能の違いはあまり大きくない．HMM の認識単位や評価に用いる単位を考慮して選ぶべきである．識別学習は，学習データ量が少ない場合や，連続 10 数字などの小規模なタスクの場合に，特に効果が大きいことが知られている．

これらの識別学習におけるパラメータ推定では，Baum-Welch 法を用いた HMM の学習で得られたパラメータ値を初期値とする．MMI 学習では，拡張 Baum-Welch（EBW）アルゴリズムという解析的な学習方法が用いられる．その他の三つでは，EBW を直接用いることができず，勾配降下法が用いられることが多い．いずれの場合も収束が保証されず，また，収束する場合も制御パラメータの値により効率が大きく異なるので，制御パラメータの注意深い設定が必要である．

また，大語彙連続音声認識においては，式 (2･1) の右辺分母を厳密に計算することは難しい．そこで，いったん初期モデルで認識して正解と競合する仮説を求め，その確率の和で代用することが多い．競合仮説として，最も確率の高い競合仮説を一つ選ぶ方法，確率の高い仮説を複数個選ぶ方法（N ベスト法），単語グラフ全体を用いる方法，などがある．

■参考文献

1) L. Rabiner and B. H. Juang, “Fundamentals of speech recognition,” Prentice Hall, pp.321-386, 1993.
2) 鹿野清宏・伊藤克亘・河原達也・武田一哉・山本幹雄, “音声認識システム,” 情報処理学会, pp.36-41, 2001.
3) X He, L. Deng, and W. Chou, “Discriminative learning in sequential pattern recognition,” IEEE Signal Processing Magazine, vol.25, no.5, pp.14-36, 2008.

■2 群 - 7 編 - 2 章

2-2 言語モデル

（執筆者：伊藤 彰則）[2009 年 4 月 受領]

2-2-1 言語モデルとは

音声認識の言語モデル（language model）は，音声信号を単語列として認識する際に，認識されるべき単語列が持っている制約を記述するモデルである．孤立単語ではなく，連続して発声された文を認識する場合に主に用いられる．数学的には，単語列 $w_1, w_2, \ldots, w_N$ に対して，それが認識単語列として許容されるかどうかの判定関数

$$\varphi(w_1, w_2, \ldots, w_N) \in \{0, 1\} \tag{2·2}$$

または，単語列の同時出現確率

$$0 \leq P(w_1, w_2, \ldots, w_N) \leq 1 \tag{2·3}$$

と表すことができる．文法に基づく言語モデルは前者の形式，統計的言語モデルは後者の形式で主に表現される．

2-2-2 文法による言語モデル

認識したい単語列をルールのかたちで表現したものが文法（grammar）である．音声認識のための文法としては，正規文法（regular grammar）またはそれと等価な有限状態オートマトン（finite state automaton）が使われることが多く，場合によっては文脈自由文法（context free grammar）も使われる．文法による言語モデルは，人手によって記述されることが多いため，主に認識対象が限定的で規模の小さい認識タスクに用いられる．

2-2-3 N グラムによる言語モデル

統計的言語モデルは，単語列の出現確率を推定するモデルである．統計的言語モデルとして最も広く用いられているのが N グラム (n-gram) モデルである．一般に式 (2·3) をそのまま推定することは困難であるため，N グラムモデルでは，ある単語の生起確率が直前の $n-1$ 単語のみに依存するという仮定を置く．すると，

$$P(w_1, \ldots, w_N) = P(w_1) \prod_{i=2}^{N} P(w_i | w_1 \ldots w_{i-1}) \tag{2·4}$$

$$\approx P(w_1) \prod_{i=2}^{N} P(w_i | w_{i-n+1} \ldots w_{i-1}) \tag{2·5}$$

となる．n としては 3 程度のものがよく用いられている．$n = 1$ のときのモデルをユニグラム（unigram），$n = 2$ のときをバイグラム（bigram），$n = 3$ のときをトライグラム（trigram）と呼ぶ．

N グラムの確率を推定する際には，大量の学習サンプル（コーパス）における単語連鎖の出現頻度を用いる．単語列 $w_1 \ldots w_k$ の出現頻度を $N(w_1 \ldots w_k)$ とすると，最尤推定によ

るバイグラム確率は

$$P_{ML}(w_i|w_{i-1}) = \frac{N(w_{i-1}w_i)}{N(w_{i-1})} \tag{2・6}$$

と表される．しかし，この計算方法では，コーパスにたまたま出現しなかった単語連鎖の出現確率が 0 になってしまう．これを防ぐために，確率の平滑化（smoothing）を行う．N グラム確率に対する代表的な平滑化手法としてバックオフ平滑化（back-off smoothing）がある[1)]．これは，出現しなかった N グラム確率を，（N-1）グラム確率によって求める手法である．バックオフ平滑化によるバイグラム確率 $P(w_i|w_{i-1})$ は，

$$P(w_i|w_{i-1}) = \begin{cases} \lambda(w_{i-1}w_i)P_{ML}(w_i|w_{i-1}) & \text{if} \quad N(w_{i-1}w_i) > 0 \\ \alpha(w_{i-1})P(w_i) & \text{else if} \quad N(w_{i-1}) > 0 \\ P(w_i) & \text{otherwise} \end{cases} \tag{2・7}$$

となる．ここで $0 < \lambda(w_{i-1}w_i) < 1$ は，コーパスに出現しなかった単語連鎖に確率を付与するために，出現した単語連鎖の確率を減らすための係数であり，ディスカウント（discount）と呼ばれる．また，$\alpha(w_{i-1})$ は，確率の総和を 1 にするための正規化係数である．ディスカウントの決定法には，Good-Turing 法（Good-Turing discounting）[1)]，Kneser-Ney 法（Kneser-Ney discounting）[2)]，Witten-Bell 法（Witten-Bell discounting）[3)]などがある．

2-2-4 最大エントロピー法による言語モデル

N グラムモデルは，ある単語に対して，直前の 1〜2 単語の制約しか反映することができない．これに対して，最大エントロピー法（Maximum-Entropy method, MaxEnt）による方法は，直前の単語だけでなく，もっと離れた場所にある単語や，文の品詞や構造など，様々な制約を確率に反映させることができる[4)]．

言語的なコンテキスト h において，ある単語 w が出現する確率 $P(w|h)$ を求めることを考える．このとき，コンテキストとしては，直前の 1 単語や直前の 2 単語だけでなく，直前の自立語と機能語，当該単語の文頭からの位置，直前の単語の品詞など様々なものを考えることができる．これらのうち，ある特定の制約 i（例えば，「直前の単語が “that” である」，など）を考え，これを制約関数 $f_i(h, w)$ のかたちで表す．このとき，

$$f_i(h, w) = \begin{cases} 1 & \text{if} \quad (h, w) \text{ が制約 } i \text{ を満たす} \\ 0 & \text{otherwise} \end{cases} \tag{2・8}$$

とする．このとき，すべての制約を考慮した条件付き確率 $P(w|h)$ は

$$P(w|h) = \frac{1}{Z(h)} \prod_i \exp(\lambda_i f_i(h, w)) \tag{2・9}$$

のように表すことができる．ここで λ_i は制約 i に関する係数であり，コーパスにおける制約の出現頻度を元に，GIS（Generalized Iterative Scaling）あるいは IIS（Improved

Iterative Scaling）アルゴリズムによって推定することができる．また，$Z(h)$ は確率の総和を 1 にするための正規化係数である．

2-2-5 トピックモデルによる言語モデル

N グラムや MaxEnt は，ある単語の確率計算に対して局所的な制約を主に反映させる．これに対して，もっと大きい文書全体の話題（トピック）を確率に反映させる枠組も用いられる．このような枠組には，潜在意味解析（Latent Semantic Analysis, LSA）[5]，確率的潜在意味解析（Probabilistic Latent Semantic Analysis, PLSA）[6]，潜在的ディリクレ配分（Latent Dirichlet Allocation, LDA）[7]などがある．これらの方法は，文書全体を空間上の点（あるいは確率分布）として表現し，それを条件として単語の出現確率を推定する．

■参考文献

1) S.M. Katz, “Estimation of probabilities from sparse data for language model component of a speech recognizer,” IEEE Trans. ASSP, vol.35, pp.400-401, 1990.
2) H. Ney, U. Essen and R. Kneser, “On structuring probabilistic dependences in stochastic language modelling,” Computer Speech and Language, vol.9, no.1, pp.1-38, 1994.
3) I.H. Witten and T.C. Bell, “The zero-frequency problem: Estimating the probabilities of novel evants in adaptive text compression,” IEEE Trans. Information Theory, vol.37, no.4, pp.1085-1094, 1991.
4) R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modelling,” Computer Speech and Language, vol.10, no.3, pp.187-228, 1996.
5) J.R. Bellegarda, “A latent semantic analysis framework for large-Span language modeling,” Proc. EUROSPEECH-1997, pp.1451-1454, 1997.
6) T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” Machine Learning, vol.42, no.1-2, pp.177-196, 1999.
7) D.M. Blei, A.Y. Ng and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol.3, pp.993-1022, 2003.

■2 群 - 7 編 - 2 章

2-3 大語彙連続音声認識アルゴリズム

（執筆者：李晃伸）[2009 年 11 月 受領]

大規模なコーパスに基づく音響モデル及び言語モデルの発展により，発話内容をあらかじめ限定しない文発声の自動書き起こし（ディクテーション）の研究が盛んに行われるようになった．一般に，数万語以上の語彙及び単語 n-gram を用いるこのタスクを，大語彙連続音声認識（Large Vocabulary Continuous Speech Recognition; LVCSR）と呼ぶ．

大語彙の連続音声認識では，出現しうる文仮説の組合せは膨大であり，すべての可能な候補についてスコアを求める（＝全探索）ことは実質的に不可能である．このため，効率良く最尤仮説を見つけることのできる解探索アルゴリズムが重要となる．また，大規模な単語 n-gram や音素環境依存 HMM を用いる場合，パラメータ数が膨大であり，そのメモリ量や尤度計算時間も，実際のシステムでは大きな問題となる．

音声認識のアルゴリズム及び計算技法はこれまで多様な手法が提案されてきた．ここでは，音声認識の基本アルゴリズムを述べた後，各種の技法をおおまかに分類して概説する．また，信頼度計算や出力形式，具体的な認識エンジンについても述べる．

2-3-1 基本アルゴリズム

統計的音声認識は，与えられた入力音声系列 X に対して $P(W|X)$ が最大，すなわち $P(X|W)P(W)$ が最大となる単語列 W を求める問題である．$P(X|W)$ が音響モデル，$P(W)$ が言語モデルから与えられる．ただし，実際には確率値のダイナミックレンジが大きく異なることから，言語モデルの確率に重みを乗じることが行われ，短い単語のわき出しを防ぐために単語ごとに負の挿入ペナルティを加えることも広く行われる．また，計算機上の尤度計算は処理速度や演算誤差を考慮して対数で行われる．これらより，ある仮説 W のスコア $f(W)$ は，言語重み LM_w と単語挿入ペナルティ LM_p，仮説の単語数 $\mathrm{len}(W)$ から以下のように計算される．

$$f(W) = \log P(X|W) + LM_w \cdot \log P(W) + LM_p \cdot \mathrm{len}(W)$$

探索アルゴリズムの使命は，この $f(W)$ が最大となる W を求めることである．具体的には，認識処理を適当な単位（フレームや単語）で進めつつ，有望な部分を推定しながら必要な精度で仮説を計算・展開していく．この有望部分の推定誤りや計算の近似誤差は，最尤仮説が得られない，いわゆる探索誤り（サーチエラー）を引き起こすため，少ない計算量で最尤解を見つけられる効率の良いアルゴリズムが求められる．

2-3-2 仮説空間の管理

認識中の仮説をどのように扱うかは探索アルゴリズムの主要課題である．仮説展開の方法は，大きく動的展開と静的展開の二つに分けられる．動的展開では，途中の時点までの照合結果を元に次に接続しうる仮説を動的に接続・展開していく．ただし，入力音声と仮説の最適なマッチングは探索途中では得られないため，常に多くの仮説を保持する必要があり，容易に幅優先に陥りやすい．このため，不要な仮説を発見して枝刈りする，類似した単語履歴

を持つ仮説を束ねて単一化する，数フレーム先を荒く照合した結果から展開仮説をあらかじめ絞り込むなどの効率化手法が必須となる．

H. Ney らは，単語仮説の履歴依存性に近似を導入することで，仮説ネットワーク全体をコンパクトなループで表現する単語対探索法を提案した[1)]．ある単語仮説内の音響尤度計算（Viterbi パス）が，直前の単語のみに依存し，それ以前の仮説履歴には依存しないという仮定のもと，全単語仮説を並列に並べてプレフィックスを共有した木構造化辞書を，直前単語ごとに個別に持つよう探索空間を構成する．全体を単純なループで効率良く表現でき，仮説空間の制御も容易であることから，広く用いられている．

近年，計算機能力の進展に伴い，言語モデル・単語辞書・音響モデルのすべての制約を単一のオートマトンネットワークとして静的に表現する，重み付き有限状態トランスデューサー（weighted finite state transducer; wFST）に基づく音声認識[2)]が広がってきた．デコーダとモデルの問題を切り離せ，多様な制約を探索に直接組み込みやすいという長所を持つ．当初はあらかじめ静的に合成・最適化する方法が用いられたが，必要な部分を動的にオンライン合成・最適化する on-the-fly デコーディング手法も提案されている．

2-3-3 探索と枝刈り

探索アルゴリズムとして，一般にフレーム同期ビーム探索が用いられる．全仮説候補を並列に扱い，入力フレームに同期して Viterbi 演算を並行に進める．仮説間の尤度比較が簡単に行える，音声入力と並行処理できリアルタイム性が高いなどの利点を持つ．そのほか，事前の照合情報が与えられるマルチパス探索では A* 探索も用いられる．

仮説展開における枝刈りの基準は，上位 N 位を残す順位基準と，最尤仮説から一定スコア以内を残すスコア基準の二つが標準的に用いられる．順位基準の発展としてヒストグラム法がよく用いられるほか，探索中に動的に幅を制御する方法なども提案されている．

2-3-4 先読み

探索において「先読み」を行うことで，仮説候補を効率良く絞り込むことができる．音響的情報に基づく絞り込みでは，数フレーム先を荒いモデルで照合した結果から展開仮説をあらかじめ絞り込む一音素先読み（1-phoneme lookahead）がよく用いられる．

また，認識時に木構造化辞書では末尾近くに達するまで単語が決定できず，言語確率の付与が遅れる．より早期に言語確率を適用するため，語頭から到達可能な単語集合の確率の最大値を漸次的に適用する LM factoring（LM pushing）が行われる．これは言語的な先読みといえる．ただし，n-gram では上記の最大値算出のために常に木の先頭で全単語の出力確率を求めることになり計算コストが高いため，キャッシュの工夫やノード間の依存関係の最適化など，様々な計算量削減方法が提案されている．また，共有ノードでは 1-gram 確率で代用すれば，静的に割当てができるので計算量を大幅に削減できる．

2-3-5 音響尤度の近似計算

典型的な大語彙連続音声認識用の音響モデルでは，数万個以上の多次元ガウス分布の出力確率計算を 1 入力フレームごとに行うことになる．実際の認識システムではこの音響尤度計算が処理時間の多くを占める．この計算量削減のための近似計算手法として，音響空間をあらかじ

めクラスタリングして入力近傍の分布のみを選択・計算するガウス分布予備選択（Gaussian selection）や，出力確率の計算過程で上位の分布のみを動的に選択するガウス分布枝刈り（Gaussian pruning ）などが用いられる．

2-3-6 信頼度

認識結果に対して信頼度（あるいは確信度）を付与することが広く行われる．仮説の事後確率 $P(W|X)$ に基づく方法が主流であり，値が 1.0 に近いほど競合候補が少ない，すなわち認識結果の信頼度が高いとする．ある W の仮説の事後確率は以下の式で与えらえる．

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} = \frac{P(X|W)P(W)}{\sum_W P(X|W)P(W)}$$

実際には，上式の分母は登場したすべての仮説の和で近似される．また，値の偏りを防ぐために確率に小さな値（< 1.0）のスムージング係数 α を乗じることが経験的に行われる．この信頼度は，語彙外発話の棄却や認識結果中の単語の取捨選択などによく用いられる．

2-3-7 複数候補の出力

実際の応用システムでは，認識結果の出力として，上位 N 位を列挙する N-best 形式のほか，単語グラフやコンフュージョンネット形式が広く用いられている．

単語グラフは，認識結果の尤度の高い単語仮説候補についてそのスコアと始終端時間を記録し，それらをつないで仮説候補集合全体をグラフで表現するものである．N-best 形式に比べてより多くの仮説をコンパクトに表現できる利点があり，多段階に候補を絞り込むマルチパス方式の音声認識システムにおいて中間表現としてよく用いられる．コンフュージョンネット[3)]は単語グラフ上の全単語をクラスタリングして時間軸方向に順列化したもので，更に仮説候補をコンパクトに表せるが，個々の単語の正確な区間情報は失われる．

2-3-8 認識エンジン

音声認識アルゴリズムを実装したソフトウェアはデコーダあるいは認識エンジンと呼ばれ，音声認識システムの中核をなす．近年，音声認識研究の発展に伴い，様々な研究・開発機関で認識エンジンの実装が進んだ．オープンソースの大語彙連続音声認識エンジンとして，国内では Julius（`http://julius.sourceforge.jp/`）が最も広く利用されている．2 パス型の認識器で，小さいプロセスサイズで効率良く探索が行える[4)]．日本語のモデルも併せて公開されているほか，一般的なモデル形式に対応しているため，様々なシステムを容易に構築できる．ほかのオープンソースエンジンとしては，Sphinx, Juicer, RWTH ASR などが挙げられる．

■参考文献

1) H. Ney and S. Ortmanns., “Progress in Dynamic Programming Search for LVCSR.,” In Proc. of the IEEE, vol.88, no.8, pp.1224-1240, 2000.
2) M. Mohri et al., “Weighted Finite-State Transducers in Speech Recognition.,” In Computer Speech and Language, vol.16, no.1, pp.69-88, 2002.

3) L. Mangu, E. Brill and A. Stolcke., “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network.,” In Computer Speech and Language, vol.14, no.4, pp. 373-400, 2000.
4) A. Lee, T. Kawahara and K.Shikano., “Julius – an Open Source Real-Time Large Vocabulary Recognition Engine–.,” In Proc. EUROSPEECH, pp.1691-1694, 2001.

■2 群 - 7 編 - 2 章

2-4 話者・環境適応

（執筆者：小坂哲夫）[2009 年 4 月 受領]

音声認識は隠れマルコフモデル（Hidden Markov Model, HMM）の使用により，飛躍的に性能が向上したが，いまだに実用的な観点からは十分な性能が得られていない．実用化するに当たって種々の問題が存在するが，その中でも，話者性の問題及び背景雑音など音響環境の問題は解決すべき重要な課題となっている．HMM に代表される確率モデルでは，学習データを収集し，それを用いてモデルパラメータの推定を行う．学習データと評価データの性質が一致する場合は高い認識性能を示すが，実際には前述した話者性や音響環境の違いなどによりミスマッチが起き認識性能が低下する．この問題を解消する手段として，適応技術が精力的に検討されている．解決すべき課題としては，1) より少ない適応データで性能向上を図る，2) 適応後の認識性能をより高める，の 2 点が挙げられる．本節では適応技術のうち話者適応及び環境適応について述べる．

2-4-1 話者適応及び環境適応の概要

音声における個人性は，調音器官の構造差を主とする生得的な要因と，発声法の違いなどによる非生得的な要因に分けられる．前者に関しては性別や年齢による差が大きく，後者については例えば方言の影響が代表的な例である．個人性と音韻性を分離できれば問題は解決するが，現状では困難である．確率モデルの利用により個人性の変動をある程度表現できるため，多数話者の音声データを用いて学習した不特定話者モデルが広く用いられている．このモデルは多数話者の平均的な特徴を表現するが，必ずしも認識対象話者の音声の特徴とは一致しない．このため特定話者モデルを用いた場合よりも性能が低下する．以上の問題を解決する手段として種々の話者適応法が提案されている．この方法では，認識対象話者の少量の適応データを用いて，モデルパラメータや入力特徴量を調整しミスマッチを低減する．

一方環境適応は，背景雑音などの雑音の影響や，電話回線の特性の影響などによる音響環境の変動に対処するための適応法である．適応による音響環境の変動への対処は，基本的には話者適応と同様な方法を利用することができる．例えばある種の雑音下で発声された音声の認識を行いたい場合，その雑音環境下の音声データを収集し適応させることにより性能向上が見込める．またその適応音声データの話者が認識対象話者と同一である場合には，環境適応と同時に話者適応も行われるという利点がある．

話者や環境に適応する場合，特徴抽出部から得られる音声の特徴量を音響モデルに合わせる方法と，音響モデルのパラメータを変更し音声の特徴量に合わせる手法に大別できる．前者は一般的に正規化法，後者はモデル適応法と呼ばれる．正規化法とモデル適応法の概念図を図 2・4 に示す．正規化法では個々の入力特徴ベクトルが何らかの変換式に従い，特徴空間上で音響モデルへ近づくように変換される．モデル適応法では逆に，モデルが入力特徴ベクトルに近づくようモデルパラメータの変更が行われる．

適応手法は適応時における教師信号の有無により，教師つき適応と教師なし適応に分類される．前者は発話内容が既知の条件で適応を行うが，後者は未知の条件のため，教師つき適応と比べ適応性能は低下する．また適応によるモデルパラメータの更新をどの時点で行うかにより，バッチ適応とオンライン適応に分類できる[1)]．前者では，適応用のデータがすべて

揃った状態で適応を行う．一方後者では，発声中随時，例えば 1 発話ごとに適応を行う．よって適応に使用できるデータが少ないため，データ量に対する工夫が必要となる．

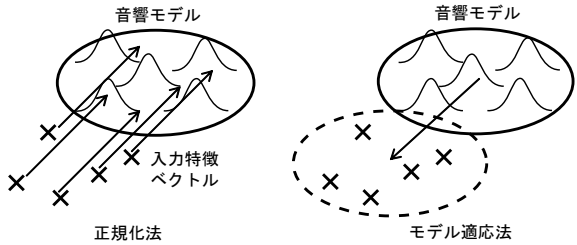

図 2・4　正規化法及びモデル適応法の概念図

2-4-2 正規化法

本節では，音声の特徴量を変更してモデルに合わせる正規化法について概説する．ケプストラム平均正規化法（Cepstral Mean Normalization, CMN）は伝送特性の差がケプストラムの差として表れる性質を利用した方法で，各フレームのケプストラムの値からケプストラムの長時間平均を差し引くことにより正規化を行う．また平均のみならず分散も正規化する手法としてケプストラム分散正規化法（Cepstral Variance Normalization, CVN）も提案されている．上記 CMN や CVN では線形の変換を行うが，ヒストグラム正規化法（Histogram Equalization, HEQ）では非線形の変換により種々の雑音に対するミスマッチを低減させることが可能である[2]．この方法では学習データと評価データのケプストラム係数の累積頻度分布が一致するような変換を行う．話者による声道長の違いを正規化する方法として，声道長正規化法（Vocal Tract Length Normalization, VTLN）が提案されている[3]．声道長は個人差が大きく，特に男女では大きな差があることが知られている．声道長が異なるとホルマント周波数も異なり，認識性能の低下を招く．そこで対象話者の声道長を求め，標準的な声道長を持つ話者のスペクトルに変換する．本手法はごく少量のデータで正規化が可能という特徴がある．

モデル適応においては，適応前の初期モデルとして一般的には不特定話者モデルが使用される．しかし話者適応のための初期モデルを考えた場合，話者間の変動を含む不特定話者モデルよりも，標準的な 1 名の話者のモデルを利用した方がよいと考えられる．これを実現する方法として話者正規化学習法（Speaker Adaptive Training, SAT）が提案されている[4]．SAT ではまず各学習話者のデータを標準話者のデータへと線形変換を行い，この変換後のデータを用いて学習を行う．作成されたモデルは話者変動が抑制されているためコンパクトモデルと呼ばれる．認識時には学習時とは逆向きの変換によりモデルを適応する．

2-4-3 モデル適応

音響モデルのパラメータを変更し適応を行うモデル適応について述べる．モデル適応はモデルの種類により適応法が異なるが，ここでは現在最も広く用いられている連続分布 HMM に関する適応について説明する．モデル学習に一般的に使用されている最尤推定も，特定の

話者のデータを適応データとして用いれば話者適応として利用できる．しかし話者適応においてはより少ないデータで適応できることが望まれる．最大事後確率推定法（Maximum A Posteriori Estimation, MAP）は事前知識を利用し，適応データ量に応じて効果的に適応をする方法である[5]．例えば連続分布 HMM の平均ベクトルの MAP 推定は以下の式で求められる．

$$\hat{\mu}_{ik} = \frac{\tau_{ik}\mu_{ik} + \sum_t c_{ikt}x_t}{\tau_{ik} + \sum_t c_{ikt}} \tag{2･10}$$

ここで x_t は時刻 t の特徴ベクトル，c_{ikt} は時刻 t の特徴ベクトルが状態 i，混合要素 k から出力される確率，μ_{ik} は初期モデルの平均ベクトルの値，$\hat{\mu}_{ik}$ はその MAP 推定値である．この式より MAP 推定値は，初期モデルの平均ベクトルと適応データの最尤推定値をデータ量に応じて補間した値として与えられることが分かる．初期モデルは一般的に不特定話者モデルが用いられる．このためデータ量が少ない場合は不特定話者モデル，データ量が多い場合は最尤推定された特定話者モデルの性能に漸近する．この方法では，適応データに存在するサブワードのモデルしか適応できないため，トライフォン HMM のようにモデル数が多い場合は大量に適応データが必要になる．

話者間の線形写像を用いてモデル適応を行う方法が種々提案されている．写像は以下のアフィン変換の式で表現され，HMM の平均ベクトル $\boldsymbol{\mu}$ が $\hat{\boldsymbol{\mu}}$ へ変換される．

$$\hat{\boldsymbol{\mu}} = A\boldsymbol{\mu} + \boldsymbol{b} \tag{2･11}$$

目的に応じて変換行列 A のみを用いる方法，バイアス $\boldsymbol{b}$ のみを用いる方法，あるいは両方を用いる方法がある．A 及び $\boldsymbol{b}$ を尤度最大化基準で推定する方法は最尤線形回帰法（Maximum Likelihood Linear Regression, MLLR）と呼ばれ，話者適応法として広く用いられている[6]．MLLR 法ではパラメータ $\{A, \boldsymbol{b}\}$ を複数のガウス分布で共有することにより，MAP 推定で問題となる適応データ不足の問題に対処している．式 (2･11) においてバイアス $\boldsymbol{b}$ のみを推定する方法として Signal Bias Removal（SBR）法が提案されており，乗算性ひずみの対策として用いられる．

2-4-4 その他の手法

多数話者の話者間の関係を用いて適応する方法が種々提案されている．話者クラスタリング法は，音声の特徴が類似した話者で話者クラスモデルを作成し，複数のクラスモデルから認識対象音声に近いモデルを選択する方法である．また固有声法（Eigenvoice）では，多数の特定話者モデルから得られた話者ベクトルから，主成分分析により固有ベクトルを得て話者適応に利用する[7]．

■参考文献

1) 篠田浩一，“確率モデルによる音声認識のための話者適応化技術，” 信学論 (D-II)，vol.J87-D-II, no.2，pp.371-386, 2004.
2) A. de la Torre, J.C. Segura, C. Benitez, A.M. Peinado, A.J. Rubio, “Non-Linear transformations of the feature space for robust speech recognition,” Proc. ICASSP2002, pp.401-404, 2002.

3) E. Eide and H. Gish, “A parametric approach to vocal tract length normalization,” Proc. ICASSP96, pp.346-348, 1996.
4) T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, “A compact model for speaker-adaptive training,” Proc. ICSLP96, pp.1137-1140, 1996.
5) C.H. Lee and J.L. Gauvain, “Speaker adaptation based on MAP estimation of HMM parameters,” Proc. ICASSP93, pp.558-561, 1993.
6) C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171-185, 1995.
7) R. Kuhn, P. Nguyen, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, “Eigenvoices for speaker adaptation,” Proc. ICSLP98, pp.1771-1774, 1998.

■2 群 - 7 編 - 2 章

2-5 雑音に頑健な認識

（執筆者：北岡教英）[2009 年 8 月 受領]

音声認識の入力となる音声波形は，様々な外乱の影響を受けている．主なものとしては，マイクロフォンに音声以外の音源から混入し波形に加法的な変形を加える加法性雑音と，残響や伝送系により加えられる乗法性の雑音がある．サンプリングされ，離散化された音声波形を $s[t]$，前者の雑音を $n[t]$，後者を $h[t]$ とすると，一般に観測される音声 $x[t]$ は

$$x[t] = h[t] * s[t] + n[t] \tag{2·12}$$

と表せる．ここでは特に断らない限り，加法性雑音 $n[t]$ を単に雑音と呼び，これに対し頑健な音声認識手法を扱う．

2-5-1 耐雑音への対処法の分類

一般に，耐雑音性を向上するための手法は大きく 3 分類されることが多い．すなわち，1. 雑音に強い特徴の抽出， 2. 雑音の抑圧，3. 音響モデルにおける対応，である．また，発声自身が連続的であるなどのために，長い音響信号中に音声が部分的に含まれる場合には，音声以外の区間の雑音が音声として認識されるなどの誤りも多くなるため，これらの手法の適応前に音声区間検出（Voice Activity Detection; VAD）が行われることが多い．

2-5-2 音声区間検出

（1）明示的な VAD

音声区間検出は古い歴史があり，当初は音声区間のみを切り出してマッチング手法により音声認識を行うことを目的に導入された．一般的には有声音のパワーと，無声子音部の検出のための零交差数（zero-crossing）への閾値処理により，フレーム単位で音声／非音声を判定する．しかし，雑音下ではこれらは性能が低い．そこで，雑音抑圧を行って，パワーやより高度な方法で検出する方法が多く研究されている．その一つは，音声のスペクトルや LPC 残差などの特徴量は高次の統計量が音声では特殊な値を取ることを利用した方法[1, 2]で，標準化された手法にも利用されている[3]．また，音声と雑音の混合正規分布などによるモデル化に基づいた識別手法も多く研究されている[4, 5]

（2）暗黙的な VAD との併用

上記のような方法でも完全な VAD は難しく，わき出し（雑音を音声と判定）や脱落（音声を雑音と判定）が生ずる．特に後者は音声認識で復旧が難しいため，一般にはある方法で得られた音声区間をやや延長したり，短い音声区間は雑音区間と修正するなどの後処理を行い，わき出しの部分は雑音区間の音響モデルを併用してマッチング段階で暗黙的に雑音と判定する方法が一般的に多く用いられ，性能もよい．

2-5-3 雑音に強い音声特徴量

特徴抽出の段階では，対数スペクトルの各コンポーネントの時系列に対するフィルタリングにより雑音に頑健にする方法がある．最もよく使われるのが RASTA フィルタ[6]

$$H_{RASTA}(z) = 0.1z^4 \times \frac{2 + z^{1} - z^{-3} - 2z^{-4}}{1 - pz^{-1}} \quad (2\cdot13)$$

で，$p = 0.98$ や $p = 0.94$ が使われる．また，自己回帰移動平均（ARMA）フィルタ[7] も有効性が示されている．前者は帯域通過，後者は低域通過フィルタに近いが，いずれも音節（4Hz 程度）や単語（2Hz 程度）をよく通過することが効果的であると考えられる．その特性を生かすため，積極的に各コンポーネントの変調周波数をクリーン音声に近づけるフィルタ設計を行う手法も提案されている[8]

2-5-4 雑音の抑圧

多くの実用システムでは現在でも古典的な手法をよく調整して用いることが多い．例えばスペクトル減算法（Spectral subtraction; SS）[9]では，観測信号 $x[t] = s[t] + n[t]$ を離散フーリエ変換して $X[k] = S[k] + N[k]$ とし，未知の雑音をなんらかの方法（一般には発声の直前の雑音）で推定した値 $\tilde{N}$ を用いて

$$|\tilde{S}[k]|^2 = |X[k]|^2 - \alpha|\tilde{N}|^2 \quad (2\cdot14)$$

として音声スペクトルを推定する．このとき，α は 1 よりやや大きめの値を用い，引き過ぎた部分は元の音声のパワーを小さくして補間する方法が良いとされる．また，推定した音声の 2 乗誤差を小さくするようにフィルタを設計するウィナーフィルタ（Wiener filter）もよく用いられる．

最近では，式 (2·12) から特徴量を得るまでの間の対数領域などで，

$$X[t] = S[t] + H[t] + \mathrm{DCT}\{\ln(1 + \exp(\mathrm{IDCT}(N[t] - H[t] - S[t])))\} \quad (2\cdot15)$$

として，クリーン音声 $S[t]$ 以外をひずみとして乗法性雑音と同時にモデル化・追従して除去する方法も研究されている[10]．

2-5-5 雑音に対応した音響モデル

最も単純で効果の高い方法は，雑音下の音声を用いて音響モデルを学習することである．クリーンな音声に雑音を重畳したデータでの学習でも大きな効果があるが，雑音下では音声の変形（ロンバード効果）があることも知られており，それらも対処可能な，実際の雑音下で収録した音声による学習が最も効果的といわれる．また，雑音が既知なら上記は効果があるが，多様な雑音下の音声で学習しておくマルチコンディション学習によって未知の雑音にもある程度対処できる．

このような重畳学習と同様の効果を音声入力環境の雑音が分かってから得るための方法として，HMM 合成法（Parallel Model Combination; PMC）[11] がある．雑音を GMM でモデル化し，クリーンな音声のモデルに正規分布上で合成することで実現する．実際にはモデルはケプストラム領域なので，逆 DCT してべき乗を取った領域で加算することになる．

また，環境・話者適応に用いられる MLLR 法や MAP 法などの適応手法も効果的である．

2-5-6 マイクロフォンアレイを用いた雑音除去

周囲雑音の混入を防ぐためには指向性マイクを用いるのが有効であるが，より積極的に指向性を得る方法としてマイクロフォンアレイがある[12)]．複数のマイクロフォンを並べて，それらへの音到来時間差を利用して音源方向を推定し，またその時間差分をずらした各マイクロフォンの信号を加算することで，その方向へ指向性を出すことができる（加算型アレイ，Delay-and-sum 型アレイ）．逆に，特定方向からの雑音があれば，その方向からの音を消去する方法もある（適応型アレイ，減算型アレイ）．

■参考文献

1) E. Nemer, R. Goubran, and S. Mahmoud, “Robust voice activity detection using higher-order statistics in the LPC residual domain,” IEEE Trans. Speech Audio Process., vol.9, no.3, pp.217-231, 2001.
2) D. Cournapeau and T. Kawahara, “Voice activity detection based on high order statistics and online EM algorithm,” IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.2854-2861, 2008.
3) ITU-T, “A silence compression scheme for G.729 optimized for terminals conforming to recommendation v.79,” annex B., 1996.
4) J. Sohn, N.S. Kim, and W. Sung, “A Statistical model-based voice activity detection,” IEEE Signal Process. Lett., vol.6, no.1, pp.1-3, 1999.
5) M. Fujimoto and K. Ishizuka, “Noise robust voice activity detection based on switching Kalman filter,” IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.468-477, 2008.
6) H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech and Audio Process., vol.2, pp.578-589, 1994.
7) C.-P. Chen and J.A. Bilmes, “MVA processing of speech features,” IEEE Trans. Audio, Speech and Language Process., vol.15, no.1, pp.257-270, 2007.
8) X. Xiao, E.S. Chng, and H. Li, “Normalization of the speech modulation spectra for robust speech recognition,” IEEE Trans. Audio, Speech and Language Process., vol.16, pp.1662-1674, 2008.
9) S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoustics, Speech and Signal Process., vol.27, no.2, pp.113-120, 1979.
10) J.C. Segura, A. de la Torre, M.C. Benitez, and A.M. Peinado, “Model-based compensation of the additive noise for continuous speech recognition experiments using AURORA II database and tasks,” Proc. Eurospeech’01, vol.1, pp.221-224, 2001.
11) M.J.F. Gales and S.J. Young, “Robust speech recognition in additive and convolutional noise using parallel model combination,” Computer Speech & Language, vol.9, no.4, pp.289-307, 1995.
12) 大賀寿郎, 金田豊, 山崎芳男, “ 音響システムとディジタル処理, ” 電子情報通信学会, 1995.

■2 群 - 7 編 - 2 章

2-6 話し言葉音声の認識・処理

（執筆者：河原達也）[2009 年 8 月 受領]

従来の音声認識の研究開発は，コンピュータ・機械やエージェント・ロボットなどとのインタフェースを主に想定して行われてきた．これに対して本節では，講演や会議などの人間が日常的に行っている音声コミュニケーションを対象とした認識・処理について述べる．このような話し言葉音声の認識ができれば，書き起こし（講演録・会議録の作成）や翻訳などの直接的な応用に加えて，今後ディジタルアーカイブとして蓄積される音声ドキュメントの要約・検索・マイニングなど様々な展開が考えられる．ただし，このような音声は，認識システムを意識しないで自然に発声されているため，音響的・言語的な変動が大きく，音声認識のうえで多くの問題がある．また，音声の忠実な書き起こしは言い淀みが多く，区分化されていないなど可読性の点で問題がある．これらの問題とその対応について述べる．

2-6-1 話し言葉音声の分類

これまでに研究の対象となっている話し言葉音声を話者数とスタイルの観点から分類したものを図 2・5 に示す．このうち，放送ニュースに関しては 1990 年代半ばから世界的に取組みが行われ，アンカー／アナウンサーの発話については 90 %を上回る認識精度を達成している．これに対して，講演や会議は一般の人が話すもので，そこまでの精度は期待できない．学会講演・口頭発表については，我が国で 2000 年代初めに大規模（660 時間）な『日本語話し言葉コーパス（CSJ）』[3]が構築され，80 %程度の認識精度を達成している．これを受けて，大学の講義を対象とした研究が世界各地で取り組まれているが，発話の自発性が高く，話題も専門的なため，おおむね 60 %程度の認識精度にとどまっている．

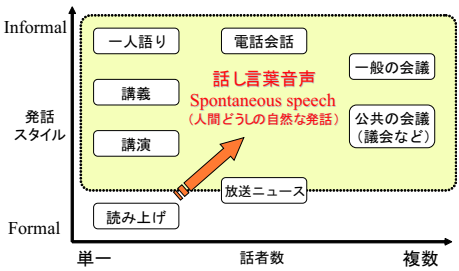

図 2・5 話し言葉音声の分類

電話会話を対象とした音声認識は，米国 DARPA プロジェクトで集中的に取り上げられ，音声の自発性が非常に高いにもかかわらず，2000 時間にも及ぶコーパスが構築された結果，80 %を上回る認識精度を実現している．一般の会議に関しては，米国 NIST の RT[4]と欧州の AMI[5]や CHIL などのプロジェクトで取り組まれ，音声認識だけでなく談話解析などの幅広い研究が行われている．公共の会議に関しては，欧州議会を対象とした TC-STAR プ

ロジェクトや我が国の国会審議を対象とした研究[6]が行われており，90 %前後の認識精度が得られている．

2-6-2 話し言葉音声の認識

話し言葉の発話に際しては，(a) 個々の音素が明瞭に発声されない（結果として音素間の分離度が小さい），(b) 発話速度が全般に速く変動も大きい，(c) 標準的な発音辞書に沿った発音がされない，(d) 口語的表現が多用され，助詞の省略など文法的でない文も多い，(e) フィラーや言い直しなどが頻発する，(f) 発話の区切りが明確でない，などの問題が生ずる．これらは，音声認識システムにおいて，音響モデル（a-b），単語辞書・発音モデル（c），言語モデル（d-e），デコーダ（f）の各々で対応が必要となり，数多くの研究が行われている．特に，VTLN に代表される音響特徴量の話者正規化や，MPE に代表される音響モデルの識別学習が大きな効果を有するという知見が得られている．

ただし，話し言葉音声認識の根源的な問題は，上記の変動・変形のパターンが図 2・5 に示した各タスクや話者によって大きく異なる反面，話し言葉を書き起こすのに膨大なコストを要するため，大規模な学習データを用意するのが困難なことである．大規模なコーパスを構築し，精密なモデルを学習すれば，十分な認識精度が得られることは，CSJ や DARPA の電話会話，国会審議の音声認識で示されているが，少しでも条件が違うと大きく性能が低下する．

したがって今後は，より汎用的なモデル化に関する基礎的な研究とともに，当該データや話者に効果的に適応する方法，更には正確な書き起こしがなくてもモデル学習を行う方法などの研究が重要になってくると考えられる．具体的に，音声認識結果を使って，モデル適応を行ったり，音響モデル学習のためのラベル作成を行ったりする研究が進められている．

2-6-3 話し言葉音声の書き起こしの後処理

話し言葉音声をそのまま忠実に書き起こしても，人間にとって読みづらいだけでなく，翻訳や要約などの自然言語処理を適用するのも困難である．その最大の要因は，フィラーや言い直しなどの言い淀みである．CSJ の学会講演では全体の 6.8 %がフィラーであり，国会の会議録作成においても 12 %程度の単語で実際の発話から修正（削除・置換・挿入）が行われている．これは，従来の枠組みで完璧な（100 %の精度の）音声認識システムを実現しても，講演録や会議録とは 10 %程度の編集距離があることを示唆している．フィラーの削除は比較的容易であるが，その他の処理はそれほど容易でない．書き起こしから整形文への変換や言い直しの検出に関してはいくつかの研究が行われている．自動整形については，統計的機械翻訳のモデルや WFST に基づく方法が提案・実装されている．言い直しの検出については，形態的特徴と韻律的特徴に基づいて CRF などの識別器を用いた手法が検討されている．

話し言葉音声の書き起こしのもう一つの問題は，文やパラグラフの区分化である．音声認識システムの出力は単純な単語の系列であり，テキストと違って，文境界（句点）やパラグラフ境界（改行）は存在しない．文境界検出に関しては，CSJ や NIST-RT[4]などで精力的に取り組まれており，前後の形態素情報とポーズなどの韻律の素性を用いて SVM や CRF などの識別器を学習することにより，正しい書き起こしではかなり高い精度（F 値で 0.9 程度）

が得られているが，音声認識結果ではおおむね認識率に応じて低下する傾向が見られる[7)]．

以上の後処理と音声認識を統合した枠組みを図 2・6 に示す．会議録や講演録は既に多く存在しているため，これらを音声認識用の言語モデルや音響モデルの学習に活用することで，前節で述べた学習データ量の問題の解決も図る．

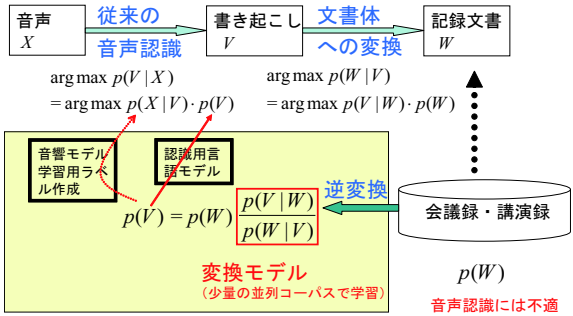

図 2・6　音声認識と後処理の統合

■参考文献

1) Sadaoki Furui and Tatsuya Kawahara, “Transcription and distillation of spontaneous speech.,” In J.Benesty, M.M. Sondhi and Y. Huang, editors, Springer Handbook on Speech Processing and Speech Communication, Springer, pp.627-651, 2008.
2) 河原達也, “〔招待論文〕 筆記録作成のための話し言葉処理技術,” 電子情報通信学会技術研究報告, SP2006-120, NLC2006-64 (SLP-64-36), 2006.
3) 前川喜久雄, “『日本語話し言葉コーパス』の概観,” http://www.kokken.go.jp/katsudo/seika/corpus/releaseinfo/040/overview.pdf, 2004.
4) Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” IEEE Trans. Audio, Speech Language Process., vol.14, no.5, pp.1526-1540, 2006.
5) S. Renals, T. Hain, and H. Bourlard, “Recognition and understanding of meetings: The AMI and AMIDA projects,” In Proc. IEEE Workshop Automatic Speech Recognition Understanding, 2007.
6) 秋田祐哉，三村正人，河原達也，“会議録作成支援のための国会審議の音声認識システム,” 電子情報通信学会技術研究報告, SP2008-99, NLC2008-44 (SLP-74-21), 2008.
7) 西光雅弘，秋田祐哉，高梨克也，尾嶋憲治，河原達也，“局所的な係り受けの情報を用いた話し言葉の節・文境界の推定,” 情報処理学会論文誌, vol.50, no.2, pp.544-552, 2009.

■2 群 - 7 編 - 2 章

2-7 話者認識・インデキシング

（執筆者：松井知子）[2009 年 4 月 受領]

話者認識[1)−9)]は，音声から本人かどうかを認証する話者照合と，話者を同定する話者識別に分けて考えることができる．近年，ブロードバンドが普及し，様々な情報サービスが展開され，信頼性，利便性が高いセキュリティがますます重要となるなか，特に話者照合への関心が高まっている．話者照合はなりすましされにくい，身体的特徴を利用するバイオメトリクス個人認証技術の一つとして位置づけることができる．音声以外の身体的特徴としては指紋，虹彩，網膜，血管，顔などが挙げられる．音声はそれらの身体的特徴と比べ，日常のコミュニケーション手段として用いられるため，自然に利用でき，また一般に普及している電話ネットワーク上で実現できるなどの利点がある．

一方，話者インデキシング[10)]では，ニュース，ボイスメール，会議などの音声について，話者の索引付けを行う．基本的に，話者認識や話者決定[11)]（どの話者がいつ発声したか“ who spoke when ”の注釈づけを行う）の方法が用いられる．インターネットなどを通して，様々な大規模な音声アーカイブが利用可能な現在，話者インデキシングの需要は大きい．

以下，話者照合，話者決定を取り上げて説明する．

2-7-1 話者照合

話者照合法は，パスワードを用いる「テキスト依存型」，特にパスワードを設定しない「テキスト独立型」，録音音声による詐称を防ぐために発声内容を毎回指定する「テキスト指定型」に大別することができる．ここでは，テキスト独立型の代表的な GMM（Gaussian mixture model）による話者照合法を例に取り，その学習と照合の手続きを説明する（図 2・7）．学習では，各ユーザと詐称者（大勢の他人）の音声を GMM を用いてモデル化する．ユーザの音声のモデル化では，種々の音韻に含まれる各ユーザの音声特徴をとらえるために，ユーザごとに数十秒から数分の発声を用いる．その発声から抽出したケプストラム係数のベクトル時系列を特徴量として用いて，そのユーザの GMM のパラメータを推定する．詐称者のモデル（universal background model; UBM と呼ばれる）のパラメータ推定には複数の話者の発声が用いられる．

照合では，学習と同じように，入力音声からケプストラム係数のベクトル時系列 X を抽出する．尤度比検定の考えに基づいて，その時系列 X とそのユーザ U，及び詐称者 I のモデルとの尤度 $P(X|U)$，$P(X|I)$ から，照合スコアとして対数尤度比 $S(X,U,I)$ を求め，あらかじめ設定したしきい値 θ と比較する．照合スコアが大きければ本人，小さければ他人であると判定する．

音声は発声環境の違い，経年変化や体調により音響的特徴量がばらつく．そのために，対数尤度比をとっても照合スコアは変動し，安定してしきい値を設定することは難しい．従来，T-norm, Z-norm などの様々な照合スコアの正規化法が検討されている．

更に音声には話者認識に有効な個人性のほか，音韻性や雑音などの様々な要因が含まれる．特徴量もしくはモデルの空間において，それらの要因を因子分析や MLLR（maximum likelihood linear regression），MAP（maximum a posteriori）などの話者適応化の手法を用いて分離して，音声に含まれる個人性を効果的に利用する方法も検討されている．

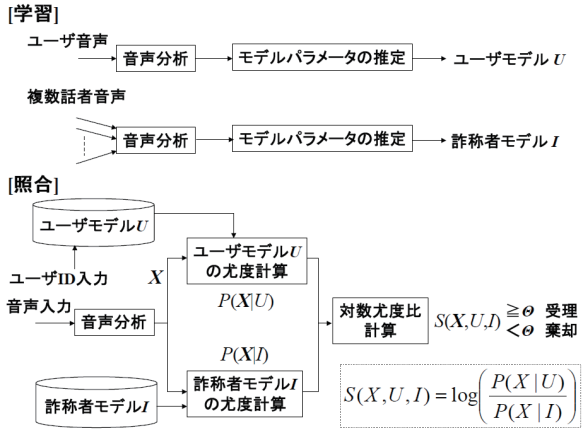

図 2・7　テキスト独立型話者照合法の学習と照合の手続きの例

2-7-2 話者決定

図 2・8 に代表的な話者決定システムの例[11]を示す．まず音声／非音声の GMM を用いるなどして“ 音声区間検出 ”を行う．次いで BIC（Bayesian information criterion）などの基準を用いて話者が交代した点を検出して，入力音声をセグメントに分ける（“ 話者交代検出 ”）．更に各セグメントについて，帯域（低域，狭帯域／電話帯域，広帯域など）や性別の大まかな分類を行ってから（“ 帯域・性別分類 ”），一つのクラスタには一人の話者だけが含まれるように各セグメントをクラスタリングする（“ クラスタ化 ”）．各クラスタをクラスタ間のクロス尤度比などに基づいて話者ごとに再結合する（“ クラスタの再結合 ”）．最後に，初期段階でのセグメント誤りを取り除くために，話者ごとに再結合されたセグメントから各話者のモデルを作成して，それらのモデルを用いて音声のセグメントをし直す（“ 再セグメント化 ”）．以上の手続きでは話者認識技術における話者のモデル化，照合スコアの正規化の方法が利用される．

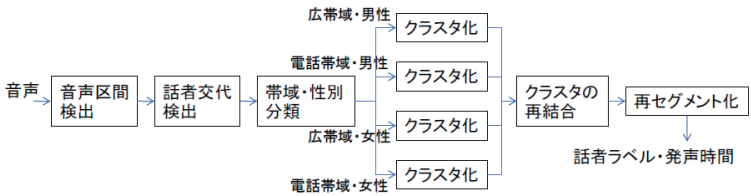

図 2・8　話者決定システムの例

■参考文献

1) 古井貞熙, “ディジタル音声処理,” 東海大学出版会, 1985.
2) S. Furui, “An overview of speaker recognition technology,” in Automatic Speech and Speaker Recognition–Advanced Topics,” Ch.2, pp.31-54, 1996.
3) J.P. Campbell, “Speaker recognition: A tutorial,” Proceedings of the IEEE, vol.85, no.9, pp.1437-1462, 1997.
4) D.A. Reynolds, “An overview of automatic speaker recognition technology,” in Proc. of ICASSP, vol.5, pp.4072-4075, 2002.
5) 松井知子, 黒岩眞吾, “音声による個人認証技術の現状と展望, −今，なすべきこと は何か! −,” 電子情報通信学会誌, vol.87, no.4, pp.314-321, 2004.
6) 松井知子, “音声による個人認証, −話者認識技術の研究動向−,” 日本音響学会誌, vol.63, no.12, pp.738-743, 2007.
7) 黒岩眞吾, 柘植覚, “話者認識技術の紹介と最近の研究動向,” バイオメトリックシステムセキュリティ研究会第 17 回研究発表会予稿集, pp.19-26, 2009.
8) D.A. Reynolds and W.M. Campbell, “Text-independent speaker recognition,” in J. Benesty, M. Sondhi , and Y. Huang (Eds.), Springer Handbook of Speech Processing, Springer, pp.763-784, 2008.
9) P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, “Speaker and Session Variability in GMM-Based. Speaker Verification,” IEEE Trans. Audio, Speech and Language Processing, vol.15, no.4, pp.1448-1460, 2007.
10) 西田昌史, 河原達也, “BIC に基づく統計的話者モデル選択による教師なし話者インデキシング,” 電子情報通信学会論文誌, vol.J87-D2, no.2, pp.504-512, 2004.
11) S.E. Tranter and D.A. Reynolds, “An overview of automatic speaker diarization systems,” IEEE TRANS. Speech Audio Process., vol.14, no.5, pp.1557-1565, 2006.