解 説 音声処理技術とその応用

4. 音声認識技術

Speech Recognition Technology by Yoshinori SAGISAKA (ATR Interpreting Telecommunications Research Laboratories).

匂坂 芳典[1]

1 ATR 音声翻訳通信研究所

1. はじめに

音声認識は人間が後天的に修得する音声言語能力を機械的にシミュレートする技術であり，多くの工学技術の発達や科学知識の増加にともない，確実に人間の能力を工学的なモデルとして取り込んで発展してきた．近年とくに，米国のARPAプロジェクト推進をはじめとしたゴールオリエンテッドな研究集約，ベンチャ企業の製品開発努力などにより，研究実用化が急速に進んできた．とりわけ，大語彙・不特定話者の連続音声認識における技術の進展には目覚ましいものがある．この研究の進展にともなって，実世界での使用をより意識した課題に研究の関心が向けられ，高騒音下の音声，自然会話音声や放送音声の認識といった挑戦的な課題についても研究が取り組まれはじめられている．本稿では，現在の不特定話者を対象とした大語彙連続音声認識システムのあらましを紹介し，そこで用いられている各要素技術の最近の進展，今後の課題について解説する．また，このような研究努力をもってしてもなお，ネイティブリスナの耳になり得ていない現在の音声認識技術を応用してゆく上での課題についても触れる．

2. 音声認識システムの構成要素

現在の不特定話者を対象とした連続音声認識の構成要素技術とそれらの間の関係を図-1に示す．この図のフローを追いながら，各構成要素技術と認識システムの動作をみてゆくことにする．

まず，入力された音声波形信号は，音声特徴抽出部で音色の違いをあらわす「音声特徴パラメータ」と呼ばれる多次元の音声特徴量の時系列に表現される．得られた音声特徴パラメータ時系列には，認識したい言語内容のほか，話者の声音や話し方の違いや入力時の環境雑音など，種々の情報が含まれている．これら多くの情報は，音声中の母音や子音といった音声言語単位が示すスペクトルの種類を多様にするため，それらの同定を難しくさせる．このため，音声言語単位の同定に不要な情報を入力パラメータ時系列から除去する，あるいは逆にシステム側でもつ単位自体の音声特徴パラメータセットを入力話者や発話環境にあわせて変形することによって，同定精度の劣化を防ぐ処理が考えられている．これらの処理は「話者・発話環境の正規化・適応」と呼ばれる．

音声言語単位の同定は，音声特徴パラメータの違いによる統計的パタン識別によって行われる．このため，各単位自体の音声特徴パラメータの統計的特性を表す数理的表現が必要である．この各単位のもつ特性の数理的表現は「音響モデル」と呼ばれ，各音声言語単位に対応する多量の音声データから，自動的に統計的モデルパラメータの形で抽出される．この統計的音響モデル作成は，音声言語単位モデルの「学習」と呼ばれる．

作成された音声言語単位の音響モデルを入力音声の音声特徴パラメータ時系列と照合することで，入力音声中に含まれる音声言語単位をみつけ出すことができるが，発話内容を一意に同定できるほど音響モデルの精度を上げることは難しい．このため，複数個の音声言語単位候補を考える必

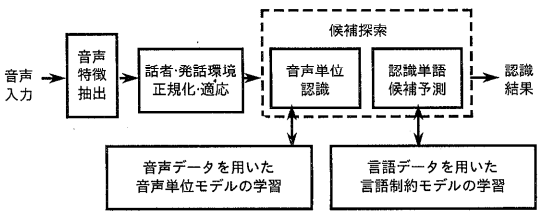


図-1 音声認識モデル構築に必要な技術

要が生ずる．この結果，入力音声長に対して指数関数的に増加する数多くの音声言語単位候補列から，効率的に正解を探索することが必要となる．

単位候補探索を効率よく進めるためには，探索アルゴリズム自体の高速化に加え，音声言語がもつ種々の性質をうまく利用して次に続く単語候補の予測を行い，探索範囲を制限することが効果的である．この探索範囲制限に用いられる音声言語がもつ制約的性質には，その言語がもつ一般的なものもあるが，限られたタスク内での使用を意図した，単語・音韻の使用頻度やそれらの連接の偏りといった認識対象タスクに依存する統計的な特徴をとらえたモデルがよく用いられる．この統計的言語制約モデルは「言語モデル」と呼ばれ，認識対象タスクに対応する言語データコーパスを用いて自動作成される．言語モデルの作成は音声言語単位の音響モデル作成の場合同様，モデルの「学習」と呼ばれる．以下，これらの音声認識構成技術の最近の進展について紹介する．

3．音声特徴抽出・認識用音響パラメータ表現

入力された音声波形そのものは，音声言語単位を特徴づけるパラメータとしては非常に冗長である．このため，音声認識では，音声の高能率伝送符号化のために考案されたスペクトル情報圧縮符号化技術による，LPC 係数やケプストラムといったパラメータが広く用いられてきた．これら音声の周波数特性を表すパラメータは，認識したい音声言語単位のスペクトル特徴を含んでいるが，より効率的で正確な音声言語単位とのパタン照合に適した特徴パラメータが必要である．

認識の目的にあった音声特徴抽出をねらい，人間の聴覚マスキング特性を反映したパラメータ化が検討されている．人間は，大きな音が入ってきた時に，その直前や直後，あるいは周波数が近い小さな音は聞こえなくなる．このマスキングは，その大きさは時間の経過とともに減少しながらもある時間継続する．このため，今聞こえている音は，今入ってきた音だけでなくそれまでに入ってきた音によって時々刻々引き起こされるマスキングを考慮することが必要である．このような聴覚特性を反映したパラメータ表現として，時間と周波数の 2 次元平面でのマスキング特性を周波数スペクトルへ畳み込む(convolution)計算方法が提案されている[1]．

一方，このような人間の聴覚特性を反映したパラメータ化とは独立に，パタン認識理論を中心に行われている音声特徴抽出の追求もある．より正確な音声言語単位の同定を行うためには，各音声言語単位固有のパタン特徴を，ほかの単位と混同せずにとらえることが重要である．音声言語単位固有のパタン特徴は各音声言語単位ごとに異なったものであり，それらを同一のパラメータで表現することは，パタンを分離する観点からは必ずしも得策ではない．このような発想のもとに，各音声言語単位ごとにそれらの同定に最も適したパタン計量を設計する方法が考案されている．このパタン分離をねらった特徴抽出では，ほかの音声言語単位との判別をよりよくするため，判別誤りを統計的に最小化する方法が用いられている[2]．

4．音声認識単位の統計的音響モデルと学習法

音声言語単位の同定のため，各音声言語単位のパタン照合を行うための「音響モデル」としては，HMM(Hidden Markov Model)が広く用いられている[3]．子音や母音を単位とした HMM では，時間によって変化するスペクトルを，3 ～ 5 個程度の状態で表される時間的に定常なスペクトルをもつ確率的信号源の遷移としてとらえる．HMM では，可能なスペクトル遷移を示す状態の個数とそれらの接続関係，入力音声特徴パラメータ x(t)に対して当該状態 i が代表するスペクトルとして観測される確率(出力確率密度関数)bi(x)，および当該状態 i からほかの状態 j への遷移確率 aij によって特徴づけられる．

これらの HMM は各音声言語単位ごとに作成され，パラメータ aij，bi(x)はそれぞれの単位に対応する音声データを用いて「自動学習」される．この自動学習ではデータとして与えられる各音声言語単位の音声特徴パラメータ系列に対して，当該音声言語単位の HMM モデルがより高い確率を示すように aij，bi(x)を逐次推定する EM (Expectation and Maximization，期待値推定・最大化)アルゴリズムが用いられる．Baum-Welch アルゴリズムはこの代表的なアルゴリズムであり，このアルゴリズムの存在が学習の自動化を可能にしている．

この HMM は，母音や子音といった音韻論上

の抽象的単位を基に，音声言語単位間の独立性を仮定した作成が考えられるが，このようなモデル作成は認識を行う上で必ずしも十分なものではない．音声スペクトルは，調音上の理由などにより，隣接する音声言語単位などによって大きく変化する．このため，隣接する音声言語単位などのコンテキストを考慮したコンテキスト依存型HMMが作成されている．コンテキスト依存型HMMを作成する場合，コンテキストの違いに対応してモデル数が増える．このため，同一の音声データでそのままモデルの学習を行う場合，モデルパラメータあたりのデータ数が減少し，モデルの推定精度が下がる．これを防ぐために，違った単位の状態間でパラメータを共有することで全体のパラメータ数を削減する方法がとられている．このパラメータを共有する状態を決めるとともに，音声言語単位ごとに異なるHMMの状態遷移関係を自動的に決定するアルゴリズムが提案されている[4]．

HMMに代表される統計的音響モデルに関する研究は10年以上にもわたって続けられており，さらに発展した音響モデルや学習法が研究されている．新たな音響モデルとしては，より精密なスペクトル変化特性を数理的に正確に記述するための工夫がなされており，セグメント・モデルと呼ばれるモデルが提案されている[4]．このモデルでは，音声を時間的に定常なスペクトルをもつ確率的信号源の遷移としてとらえるHMMに代わって，ある時間幅でのスペクトルの推移パタンとして統計的にモデル化している．このようにある時間幅をもつ動的なパタンを対象とすることで，マルコフモデルが宿命的にもつ，隣接フレーム間データの独立性仮定に起因する性能劣化を防ぐことが期待されている．

また，ニューラルネットを用いたモデルも同様な理由で研究が続けられている[4]．ニューラルネットを用いたモデルは識別モデルとして理論的に簡潔である上に，少ないパラメータで効率のよい表現が可能で，ほかの音声言語単位モデルとの識別がモデル学習時に自動的に組み込まれる利点がある．このようなニューラルネットのもつモデルの識別能力を積極的に学習に取り込み，一般化し，ほかのモデルとの認識誤りを最小化する基準に基づいた学習方法が提案されている[5]．

5．不特定話者音声認識と話者適応技術

話者や発話環境の違いによって変わる入力音声スペクトル形状の違いは，音声言語単位のスペクトルパタンとの近さを基にパタン照合を行う音声認識にとって深刻な問題である．不特定話者を対象とする音声認識では，何らかの形で発話者個人の差異を吸収することが求められる．このため，音響モデルの作成には大量の話者による音声データが用いられ，認識時にも男性と女性のように音響的特性が大きく異なるものについては，複数個のモデルが用いられる．また一方，入力音声の最初の少量データなどを用いて，認識システムがもつ音声言語単位モデルを入力話者個人のスペクトル特性を反映した音声言語単位モデルに変形する「話者適応」の方法も広く用いられている．以下のような課題が研究として取り組まれている．

(1) システムがもつ音響モデルと発話者個人の音声データによる適応とのバランス

発話者個人の音声データが少量しか使えないため，音声データ中にみられない音声言語単位も存在する．また，存在してもあまり多くのサンプル数は期待できない．このため，音響的に近い音声データを用いた補間や平滑化によってモデルを統計的に推定する必要がある．また，少量の発話者音声データに対して，システムがもつ音響モデルは豊富な音声データを用いてモデルを学習できるため，モデルを変形する上では，このデータ数の違いによる統計的な信頼度を考慮する必要がある．さらに，発話者個人の音声入力データ量が増すにつれて，信頼度も変わってくるため，データ量に応じた適応が求められる[6]．

(2) システムがもつ適応前の音響モデルの作成法

たとえば男女声の相違のように，適応前の音声言語単位モデルが発話者のスペクトル特性と大きく違う場合，適応は難しく，適応による効果は小さい．発話者に近い話者からの適応が最も効果的なので，話者を木状に階層的にグループ化して複数の音声言語単位モデルを用いる認識手法が提案されている．このような認識手法では，最初の入力音声を使って最も近い話者グループを選択し，それに対応する音響モデルを適応に用いる[7]．

(3) 動的逐次適応

話者への音響モデルの適応は通常バッチモードで行われるが，認識システムの使い勝手からすると，オンライン型の適応方法が望ましい．ベイズ適応法を用いたオンライン逐次型の処理が考案されており，適応に用いる学習データに存在しない音声言語単位のモデルも含め，ベイズ学習される[8]．

以上で紹介した音響モデルの選択・適応法に加え，音響パラメータレベルで話者固有の特徴を吸収するため，スペクトル周波数軸の非線形収縮を行う声道長正規化処理なども考慮されている．また，これら話者による入力音声パタンの相違への対処策は，環境などほかの要因による相違への対処にも用いられている．

6. 音声認識のための言語情報の利用

音声認識において，音響モデルの精度は年々向上してきているが，発話内容を一意に同定できるほど音響モデルの精度を上げることは難しく，何らかの言語的知識を制約として最終決定を行うことが必要である．この言語的な制約としては，音素・音節の種類，連接特性(phonotactic constraints)，構造などにみられる音声言語としての制約と，文節・句・節といった単位のもつ文法的，意味的な制約がある．本来，言語そのものがもつこれらの一般的な制約だけを用いて音声言語単位の同定をしてゆくのが望ましいが，現状の音響モデルの性能はまだ低く，それら一般的な言語制約だけでは緩すぎて，正解を絞り込めない．また，理想的な一般的制約を実際に得ることも難しい．

とくに，音声認識では仮名漢字変換のように入力する文字個数が既知であるわけでもなく，音声言語単位の時間長変動を許した膨大な探索をしなければならないため，候補数と入力音声長の増加にともなって照合回数が天文学的に増大する．一単語の認識だけを対象とする孤立単語音声認識では，総語数が少なければ総あたりで照合することも可能だが，連続発話を対象とした音声認識では，単語間の境界が一意的に決定できないために複数個の候補が生じ，すべての単語列を総あたり的に照合することは時間的に難しい．このため，認識対象のタスクを設定することで，そのタスクに固有な単語列の生起頻度統計を用いた言語制約が考えられている．

言語制約の統計的モデルとしては，単語のN-gramが広く用いられている．このモデルでは，N単語の連接の偏りを利用しようとするもので，あらかじめ大量の言語データベースを用いて単語間の遷移確率を計算しておく．認識時には，単語候補の音響的なパタンの近さにこの確率を加味して探索を行う．単語連接の個数Nの増加とともに正解の確率が高くなることが期待できるが，それらの確率値の推定に用いる言語データのサンプル数が少なくなるため信頼度が低下するため，実際はN＝2，3として用いられることが多い．また，出現サンプルのばらつきに対処するための数々の平滑化・補間手法が工夫されている．さらに，Nを固定せず，データの分布に従って効率的に統計的信頼度の高い単語連鎖制約を得る可変長N-gramなども提案されている[9]．

7. 単語探索戦略と統計的発音辞書

連続音声認識においては，入力音声に対し，連接可能な単語を言語的制約によって絞り込んだ探索を行い，音響モデルが与える音響的尤度(acoustic likelihood)と統計的言語制約が与える言語尤度(linguistic likelihood)に基づいて単語の同定を行う．しかし，統計的言語制約を用いても，なお認識時の入力音声と認識候補との照合回数は多い．認識が進むにつれて組合せ的に増大する単語候補列の情報を効率よく蓄え，高速に比較するための探索法が必要である．とくに，自然会話音声のように，書き言葉に比べて言語的制約が比較的自由であり，音響的にも変動が大きい音声の認識を対象とする場合には，単語仮説数は増大するため，探索は重要な問題である．

初期の認識システムでは，トップダウン的に言語制約を利かし，入力順に局所的な判断を基に単語を順次同定してゆく探索方法がとられたが，一般的に，最終決定は音声全体を眺めた上で行う方がよい結果が得られるため，多段階の探索方法が広く用いられるようになった．多段階の探索を効率よく行うためには，単語仮説数の減少が必須であり，単語グラフによる表現が広く用いられている[10]．単語グラフは，各単語の開始・終了時刻，先行単語情報，音響・言語尤度などの情報からなる文仮説をネットワーク表現したもので，多くの文仮説をコンパクトに表現することができる．単

語グラフ生成時の文仮説数削減を図るため，同音語の言語尤度の共有化，開始時刻による先行単語のマージなどが提案され，数千語規模の準実時間連続音声認識の可能性が示されている[10]．

単語の探索時には，単語の発音に基づいて音響モデルが用いられるが，実際の発音にはゆれがある．音声学的な知見をもつ専門家が人手で発音辞書を作成して，このゆれを吸収することも考えられるが，用いる音響モデルの特性にあわせた発音辞書を作り上げるのは至難の業である．1つの単語に対して，統計的に妥当な複数の発音を与えるため，統計的発音辞書の作成法が研究されている．統計的発音辞書としては，学習データ中の発声数が多い単語に対しては高い統計的信頼性を保ち，データ中にほとんど現われない単語に対しても，音響モデルの認識特性を反映した発音記号を与えるものが期待される．辞書作成の一方法として，ニューラルネットを用いた複数発音の統計的推定モデル化がなされている[11]．この方法を用いた認識実験結果が示すように，複数発音を用いることにより，認識率が向上するばかりでなく，認識のあいまいさが減少するため，認識にかかる時間の短縮も期待できる．

8. おわりに

以上で紹介したように，音声認識では，発話者や話し方，発話環境によってばらつきの大きい音声信号に対して，音声的な知識や言語的な知識を用い，効率よくパタン同定を行うためのモデルを考え，それらを学習することが進められている．これらのモデルを，人間が行っていると考えられる処理や人間がもっているさらに優れた能力と比較した場合，技術はまだとても人間に及びもつかない．実際，話し手や話し方，発話環境が変化しても認識能力を低下させない，いわゆるロバスト性の問題は音声認識研究の最大の研究目標になっている．また，多くの単語数を許した場合の認識能力も十分なものではなく，研究努力が続けられている．

音声認識の実用化には，ここで紹介した各々の技術の発展がさらに必要であることはいうまでもない．しかし，それにも増して，我々人間の音声言語知覚・理解機構のモデル化，コミュニケーションにおける音声言語の使われ方の数理的モデル（いわゆる対話モデル）といった音声認識技術と共用される技術の重要性がより強く認識されはじめている．世界中で進められている特定タスクへの限定的な認識技術の応用例に多くみられるように，現在の認識技術にほかの知識や技術を加えることで現実的な応用システムイメージを描けるものも少なくない．

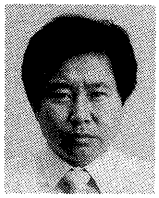
米国ARPAプロジェクト主導の種々のタスク，欧州各国の研究組織で進められている列車予約・情報案内といった自動音声応答サービスをはじめ，自動音声翻訳をねらったC-starやVerbmobileといった研究コンソーシアムでも，タスクのもつ制約をうまく利用することによるシステム化が検討され始めている．これらの例が示すように，今後も音声処理と言語処理の統合がさらに進むことが予想される．これらのシステムで用いられている音声認識技術は，確かにネイティブリスナの能力より劣るものの，たとえば外国語の音声認識といった側面からの人間の能力を考えると，少なくとも表面上は，はるかに人間の認識能力を超えた能力を示すのもまた事実である．人間と同じ能力をもつ音声認識・学習機能の基本的な追及とともに，それら不完全な認識機能の欠点をカバーしながらシステムとしてまとめあげてゆく技術が求められている．

参考文献

1) 相川清明，河原英紀，東倉洋一：順行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を用いた音韻認識，電子情報通信学会論文誌A，Vol.J76，No.11，pp.1514-1521（1993）．
2) Watanabe, H., Yamaguchi, T. and Katagiri, S.: Discriminative Metric Design for Pattern Recognition, Proc. IEEE ICASSP '95, pp.3439-3442 (1995).
3) Rabiner, L. R. and Juang, B. H.: Fundamentals of Speech Recognition, Prentice Hall (1993).
4) Singer, H., Fukada, T. and Sagisaka, Y.: Acoustic Models for Speech Recognition: A Survay, 日本音響学会春季講演論文集，pp.27-30 (1997).
5) Juang, B. H. and Katagiri, S.: Discriminative Learning for Minimum Error Classification, IEEE Trans. Signal Processing, Vol.40, No.12, pp.3043-3054 (1992).
6) Tonomura, M., Kosaka, T. and Matsunaga, S.: Speaker Adaptation Based on Transfer Vector Field Using Maximum a Posteriori Probability Estimation, Computer Speech and Language,

Vol.10, No.2, pp.117-132 (1996).
7) 小坂哲夫，松永昭一，嵯峨山茂樹：木構造クラスタリングを用いた話者適応，電子情報通信学会論文誌 D，Vol.J78-DII，No.1，pp.1-9 (1995).
8) Huo, Q. and Lee, C. H.: On-line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate, IEEE Trans. on Speech and Audio Processing, Vol.5, No.2, pp.161-172 (1997).
9) 清水 徹，山本博史，政瀧浩和，松永昭一，匂坂芳典：大語い連続音声認識のための単語仮説数削減，電子情報通信学会論文誌 D，Vol.J79-DII，No.12，pp.2117-2124 (1996).
10) 政瀧浩和，松永昭一，匂坂芳典：連続音声認識のための可変長連鎖統計言語モデル，電子情報通信学会，技術研究報告，SP95-73，pp.1-6 (1995).
11) 深田俊明，匂坂芳典：発音ネットワークに基づく発音辞書の自動生成，電子情報通信学会，技術研究報告，SP96-71，pp.15-22 (1996).

(平成9年8月4日受付)



匂坂 芳典

昭和48年早稲田大学理工学部物理学科卒業．昭和50年同大学院修士課程修了．同年日本電信電話公社（現，NTT）武蔵野電気通信研究所入社．昭和61年より国際電気通信基礎技術研究所（ATR）に出向．現在，ATR音声翻訳通信研究所第1研究室室長．工学博士．音声合成・音声認識を中心とした，音声情報処理，言語情報処理の研究に従事．電子情報通信学会，日本音響学会，IEEE，米国音響学会各会員．
e-mail:sagisaka@itl.atr.co.jp