21世紀に向けての音声認識

牧野 正三
東北大学大型計算機センター／大学院情報科学研究科

【音声認識の最近の歴史】

現在音声認識・理解システム（以下では対話システムも含める）の基本モデルは，音響処理部ではHMM（Hidden Markov Model：隠れマルコフモデル），言語処理部では単語トライグラムである．HMMは図-1に示す構造をしており，各ノードに出力確率密度関数として複数個の正規分布で表現される混合連続分布を対応させ，弧には遷移確率を対応させることが多い．1つのHMMが1つの音素や単語を表す．入力された音声に対して，通常10msごと（フレーム）に複数のパラメータ（ベクトル）が計算される．得られたベクトル時系列に対して，各HMMはその系列を出力する尤度を遷移確率と出力確率密度関数を用いて計算する．最大の尤度を与えるHMMに対応する単語や文を認識結果とする．状態の数は音素で3から7であり，単語や文は音素HMMの連結で表すことが多い．また，単語トライグラムは単語の3つ組連鎖で2次のマルコフ過程に相当する．

1980年代前半から米国や欧州で種々の音声認識・理解プロジェクトが開始され，それとともに音声コーパスの整備（数百人の話者が発声した数百文の連続音声データ）が進められた．音声データの量が増えるとともにHMMの持つ柔軟性と理論的明快さが輝きを増し，1980年代後半からHMMが主流となり，現在のように頑健な音声認識システムを構築することが大変容易になった．日本においても，1980年代中頃から発足したATRは音声コーパスの重要性に着目して音声データ量の増大に力を注ぎ，HMMやニューラルネットを利用した音声認識・理解システムの研究を行う基盤が整備された．現在では情報関連企業から種々の音声認識ソフトウェアや音声応用機器が製品として販売されるようになっている．

【音声認識・理解のメカニズム】

対話システムも含めた，現在の音声認識・理解システムの構成を図-2に示す．点線部分はオプションを意味する．現在情報処理振興事業協会の研究開発事業（代表：鹿野清宏奈良先端大教授）として，日本語ディクテーション基本ソフトウェアJuliusが整備されており，そこで採用されているアルゴリズムが現時点での音声認識の代表的アルゴリズムといえる．本稿では，誌面の都合上，多くの参考文献をあげるのは不可能なので，詳細は情報処理学会[1),5)]，電子情報通信学会[2)]，日本音響学会[3),4)]の論文特集や解説を参照されたい．

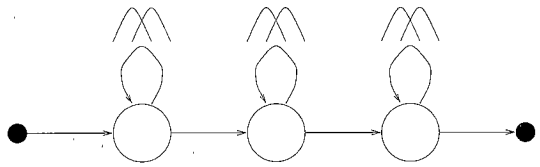

図-1　Hidden Markov Model

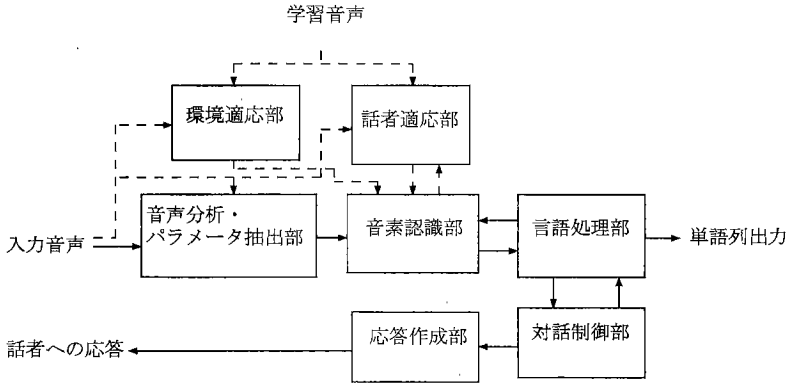

図-2　音声認識・理解システムの構成

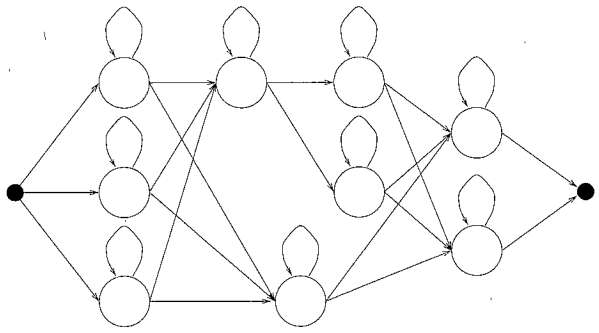

図-3　HMnet

• 音声分析・パラメータ抽出部

音声分析では，線形予測法に基づく方法を用いることが多い．特に周波数軸を聴覚心理尺度であるメルに，振幅軸を対数にとった場合のスペクトルを，コサイン変換したパラメータと近似的に等価なLPCメルケプストラムがよく利用されている．

• 環境適応部

話者のいる音環境情報を取り入れ，雑音や騒音への適応を行う．雑音としては，駅構内のような環境雑音である加法性雑音と，室内音響特性や回線特性のような乗法性雑音がある．

• 話者適応部

主に話者の声道長（声帯から唇までの長さ）に起因する音声の個人性を吸収する．発声者の声道長を推定する方法，少数の学習サンプルを用いて標準パターンを発声者に適応する方法，多数の話者の中から似た話者を選択する方法がある．

• 音素認識部

認識単位としては，言語処理部との整合性を考慮して音素や音節等が用いられることが多い．認識単位のモデル化法には，ほとんどの場合HMMが用いられている．音素の場合は，前後の音素の影響を考慮したコンテキスト依存音素HMMや，図-3に示すようなHMnet（Hidden Markov Network：隠れマルコフ網）が利用されている．現在は，言語処理部から生成された文モデルに従って音素モデルや音節モデルを連結し，直接音声パラメータの時系列とマッチングすることが多い．

• 言語処理部

文モデルは単語トライグラムを用いて生成することが多い．生成される文モデルの数は数百万以上の非常に大

きな数になるので，可能性の高い複数の文モデルのみを残し，可能性の低い文モデルは捨てていくことによって計算量を大幅に削減し，実時間化を実現している．

「Via Voice」や「Smart Voice」のような市販されているディクテーションシステムは，意味情報を利用しないで単語列をそのまま出力する．一方，音声理解システムや対話システムでは対話制御部と連動して意味理解を行う．

・対話制御部

言語処理部と連動して入力音声に対する意味理解を行い，話し手の意図を理解し行動を起こす．あるいは意味が不明な場合には話し手に質問を行うなどする．話し手とシステムの対話の過程は，あらかじめシステム内に有限状態オートマトンや意味フレームの形で蓄えられていることが多い．

対話システムでは，さらに対話に多くみられる語順の違いや冗長語に対応するため，キーワードスポッティングに基づくシステムが多い．

【未来展望】

今後の音声認識・理解システムの適用分野を考えていく上で重要なのは，インターネットや携帯電話，モバイルコンピューティングとの関係である．これらの情報基盤との競合や共存を常に考慮していく必要がある．今後の適用分野として下記の分野が考えられる．また，今後は音声だけではなく，画像，ジェスチャ，表情等の認識も含めた統合的なヒューマンインタフェースシステムを構築していく必要がある．

- 対話型介護ロボット，案内ロボット，オフィスロボット，うなずき・あいづちロボット，ペットロボット等
- 各種案内・予約システム
- 音声通訳システム
- 外国語教育システム
- ナビゲーションシステム
- 携帯電話用ダイアリングシステム
- 大規模ディクテーションシステム
- 定型文ディクテーションシステム
- エンターテインメントシステム
- 音声話題分類システム
- 音声情報検索システム
- 音声要約システム
- その他

【新しい音声認識技術】

上記システムを実現していくための，今後の音声認識技術として重要なものや興味深いものを下記に述べる．

・音声コーパス

過去10年間の音声認識・理解研究の成功は大量の音声データに負うところが大きく，その重要性はますます増大している．規模としては数千人規模の話者の発声または発話を収録することが要求されよう．また，既存データを基に，発話様式の異なるものや発話環境の異なるものを自動的に作成する方式の開発，複数データベースを結合するためのフォーマット変換方式の開発等が必要である．

・音声分析・パラメータ抽出部

現在主に利用されているLPCメルケプストラムを超える音声分析法やパラメータが模索されている．特に聴覚機構に着目されているが，大量データに基づくHMMのような統計的手法ではその差が際立ってはいない．今後の音声分析法としてはフィルタバンクやFFTで精細な分析をした後，聴覚機構を参考にした，雑音や基本周波数の影響が少ない分析法（たとえば梶田ら[6]のSBCOR法）が望まれる．パラメータとしては従来あまり利用されなかった周波数情報の利用が考えられる．たとえば菅村ら[7]の提案したLSP等の音声認識への利用も興味深い．

・環境適応部

現在の音声認識・理解システムの成功の要因の1つに環境適応技術の進展があげられる．今後は特に雑音の時間的変動を考慮した方法の開発が重要である．方法としては雑音をHMMでモデル化し，音素HMMと融合させて雑音のある音声を生成する，HMMの合成・分解に基づく方法が有効であろう．

・話者適応部

話者適応に関する研究では，日本は量的な面と質的な面の両面で優れている．松本，古井，中川らによって行われており，現在よく利用されている声道長正規化や，MLLRと同じ発想の研究が数多く行われている．たとえば，松本らは周波数軸DPに基づく声道長正規化，古井らや中川らはMLLRと同じ考えの回帰分析に基づく話者適応の研究を行っている．今後の話者適応は，篠田ら[8]が提案したように学習サンプル量を考慮して適応方法を自動的に変更していくとともに，認識結果を利用したオンラインリアルタイム話者適応に推移していくものと考えられるが，前述の先駆者の研究を統計的手法の立場から見直すことも有意義ではないかと考えている．

・音素認識部

中川は音素認識の1～2％程度の精度向上が言語処理部におけるパープレキシティの20％程度の減少よりシ

ステム全体の精度向上に有効であると予測している．音素認識の精度向上の方法としては，以下の2つがあげられる．

(a) セグメント情報の利用　今まではケプストラムの時間差分量であるデルタケプストラムが主に利用されてきたが，大量データの利用が可能になれば時間-周波数間の相関を考慮した統計的なセグメント情報の利用が進むものと考えられる．

(b) マッチング方法の再検討　これまで継続時間情報を考慮した非線型なマッチング方法が用いられてきたが，複数パターンと線形マッチングの併用に進む可能性がある．たとえば，子音から母音への遷移区間での時間伸縮は少ないのでむしろ線形マッチングを行った方がよいと考えられる．したがって，今後は区間の性質によって線形と非線型のマッチングを使い分ける方式が有望である．さらに進めて前述のセグメント情報との整合性を考えると，池田ら[9]が提案しているように，大量の音声データを利用することができれば，音素や音節ごとに，さらに各継続時間長ごとに標準パターンを用意しておき，すべて線形マッチングで認識するということも考えられる．

• 言語処理部

単語トライグラムが提案された1970年代初期の頃は，より制約の強い有限オートマトンや文脈自由文法の方が意味処理との整合を考えても有望であるというのが学会の大勢であった．しかし，ここでも大量データによる単語トライグラムの推定が威力を発揮した．特に単語トライグラムの強力な平滑化法である削除補間法やBack-offスムージングにより，トライグラム確率を精度よく推定できるようになり，未知サンプルに対して威力を発揮した．一方，有限オートマトンや文脈自由文法で実用的な規模のものを自動的に作成することは困難であり，かつ未知サンプルに対する頑健さも十分なものとはいえない．今後の方向としては，未知語を考慮した有限オートマトンや文脈自由文法の実用的な自動獲得アルゴリズムとその平滑化法の開発，単語トライグラムと有限オートマトンや文脈自由文法との融合などがあげられる．

• 対話制御部

現在よく利用されているキーワードスポッティングに基づく対話システムは自然性は高いが，大規模システムには向かない．また，冗長語等を考慮した文法を利用した対話システムは大規模システムには向くが，文法の構築，自然性に問題が残っている．今後研究を進めていくためには，まず大規模対話データベースの構築が必要である．

【これからの音声研究者へ】

今後は，学術研究機関のみならず，民間企業も含めたコンソーシアムあるいは機構を設立し，実用を想定したタスクを設定し，協力あるいは競争してデータベース整備や研究・技術開発を推し進めていく必要がある．そのためには，日本でも米国で行われているような競争型プロジェクトシステムを立ち上げるべきであろう．日本においてもディクテーションシステムでは製品が出てきているが，今後は音声対話の実用的なシステムを早急に構築する必要がある．またディクテーションやキーワードスポッティングにおいては，環境適応技術の開発をより積極的に推し進めていく必要がある．どのような環境でも音声認識・理解や対話ができることを示すことは大変重要である．また，これまで利用されていない情報，たとえば従来利用されていた母音の「あ」であるという情報のみならず，「あ」でないという情報（補空間情報）も併用していくことも必要であろう．

参考文献

1) 音声言語情報処理の現状と研究課題, 情報処理学会誌, Vol.36, No.11, pp.1011-1053 (Nov. 1995).
2) 音声言語によるコミュニケーションシステムの実現に向けて, 電子情報通信学会論文誌, J79-D-II, pp.2003-2206 (1996).
3) 音響信号処理による音声認識性能の改善, 日本音響学会誌, 53, pp.864-894 (1997).
4) 音声対話システムの実力と課題, 音響学会誌, 54, pp.783-821 (1998).
5) 音声言語情報処理, 情報処理学会論文誌, Vol.40, No.4, pp.1355-1498 (Apr. 1999).
6) Kajita, S. and Itakura, F.: Speech Analysis and Speech Recognition Using Subband-Autocorrelation Analysis, Journal of Acoustical Society of Japan (E), Vol.15, pp.329-338 (1994).
7) 菅村, 板倉: 線スペクトル対 (LSP) 音声分析合成方式による音声情報圧縮, 電子通信学会論文誌, J64-A, pp.599-606 (1981).
8) 篠田, 渡辺: 音声認識における自律的なモデル複雑度制御を用いた話者適応化, 電子情報通信学会論文誌, J79-D-II, pp.2054-2061 (1996).
9) 池田, 梶田, 武田, 板倉: 長さごとに用意されたセグメント標準パターンとの照合に基づく音声認識, 電子情報通信学会論文誌, J82-D-II, pp.308-310 (1999).

(平成11年11月30日受付)