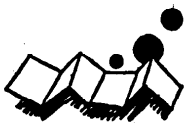
解　説



音声認識・理解研究の動向†

坪井宏之††　浮田輝彦†††

1. まえがき

人間の発声した音声を機械で自動認識する音声認識の研究は，古くから行われており，あらかじめ利用者の声をシステムが訓練して利用する特定話者用の単語認識装置などが実用化されている．音声認識の究極の目標は，話者を限定せずに（不特定話者）日常の連続的な会話音声を認識し，その内容を理解することといえる．しかし現状の技術レベルは，この目標からは大きなギャップがあるというのが実情である．

最近では，1970年代初期に行われた米国ARPAの音声理解システムの研究の後，ふたたび大規模な研究プロジェクトが開始されている．本稿では，技術的な立場から音声認識・理解研究の動向を概観する．まず，音声認識・理解の技術項目を整理した後に，研究の現状を概観する．さらに解決すべき問題点を考察し，今後の音声認識・理解技術を展望する．なお，音声情報処理全般の現状については，文献91)～100)の入門書・特集記事を参照されたい．

2. 音声認識技術

音声はマンマシンインタフェースの基本的手段の一つであるが，十分に利用されるためには，発声の制限ができるだけないことが望ましい．特に音声入力の利点の一つである入力速度を生かすためには，文などを連続的に発声した連続音声の認識が必要となる．連続音声の認識では，記号化の最小単位を音素程度とする場合と，単語にする場合が考えられるが，ここでは前者の場合について考える．これは，現在多く研究されている大語彙単語音声認識にも共通した基盤技術である．

2.1 技術項目

連続音声認識の技術項目は表-1のようにまとめられる．雑音については別の機会にゆずり，その他の項目についてみてみる．

（1）音響分析

高品質の音声の入力では，0～8000 Hz 程度の周波数範囲にわたって分析することが多い．分析区間（フレーム）長が 10～30 msec 程度，分析周期が 4～16 msec 程度ごとに，特徴パラメータを抽出することが多い．ノンパラメトリックな分析ではエネルギ，バンドパスフィルタ出力や FFT 分析などがある．周波数軸上は log スケールで等間隔にする場合や，さらに聴覚の特性を取り入れた臨界帯域(critical band)[1]で，15～30 個程度にすることが多い．パラメトリックな分析として，音源（声帯）によって駆動される音響管（口腔）をモデルとした全極モデルに基づく線形予測（LPC）係数[2),3)]や偏自己相関（PARCOR）係数[4),5)]，LPC ケプストラム係数[2),3)]などがある．

（2）音韻認識

特徴パラメータ系列を記号列に変換するときの単位をここでは認識単位という．言語学の音韻論的単位としては音素があるが，音響特徴が異なる（たとえば，“サ”と“シ”の/s/の部分の物理的な特徴は異なる）ため，より物理波形レベルに近い音声学的な記号まで含めたもの（ここではこれを音韻と呼ぶ）を用いることが多い．さらに二つの音素を組み合わせた diphone，母音核を中心とした音節（syllable），母音-子音-母音連鎖などがある．これらの単位へのセグメンテーションは，無音・有音，無声・有声などの大分類を零交差数，エネルギ，一次の LPC 係数などを用いて行い，さらにエネルギや特徴量の時間的変化により細分化することが多い．すなわち，1)パワーやスペクトル変化パラメータの極値を用いるもの[6)]，2)入力データ区分の分散を最小化するもの[7)]，3)音韻の標準パターンとの距離を利用するもの[8)]，などがある．

音韻認識は周波数領域で行われることが多く，情報

† Research Trends in Speech Recognition and Understanding by Hiroyuki TSUBOI (JAPAN ELECTRONIC DICTIONARY RESEARCH INSTITUTE, LTD. (EDR)) and Teruhiko UKITA (R & D Center, Toshiba Corp.).

†† (株)日本電子化辞書研究所
本内容は筆者が(株)東芝在職中に行った研究成果の一部である．
††† (株)東芝総合研究所

表-1 連続音声認識の技術項目

技術項目	内容	主要な課題
音響分析	音声からの特徴パラメータ抽出	音韻不変な特徴の抽出，聴覚特性の考慮
音韻認識	音韻単位への区分化とその識別	音韻の種類，動的特徴の利用，識別方式，音形規則の利用
単語認識	音韻系列中の単語の同定	同定方式とスコアリング，単語辞書の表現
話者適応・正規化処理	発声器官形状や話し方の違い	正規化処理，話者への適応化方式
雑音	騒音，伝送歪み，せき払い，紙めくり音など	音声区間の検出，認識対象語と不要語の区別

の不適切な切り捨てを少なくするために，複数候補にそれぞれスコアをつけ，セグメントラティスまたは音韻ラティスという形で結果を出力する．距離尺度としては各特徴パラメータのユークリッド距離，LPC スペクトルの対数包絡の距離である LPC ケプストラム距離，スペクトルの山部に重みを置いた WLR (weighted likelihood ratio: 重み付き尤度比) 尺度[9]，訓練サンプル内の特徴パラメータがもつ変動を直交空間で表現する複合類似度[10]などがある．最近ではスペクトルの山部の位置を重視した平滑化群遅延スペクトル距離尺度[11]なども研究されている．

識別方法はベイズ識別に基づく統計的決定理論，k-NN 識別法[12] (k-nearest neighbor classification)，単純パターンマッチング法，複合類似度法などに分類できる．それぞれの方式では識別する各カテゴリの訓練パターンを基にカテゴリを代表する内部表現を作る．ベイズ識別方式は内部表現として各カテゴリの出現確率とカテゴリに属する訓練パターンの確率密度をもつ．また，識別は訓練パターンの誤識別率を最小にする意味で最適である．k-NN 識別法の内部表現である標準パターンは，代表的な訓練パターンそのものである．各カテゴリの標準パターンすべてから入力パターンに近い k 個を選び，最も多数のパターンをもつカテゴリを結果とする方式である．また，標準パターン数と k が大きくなるにつれて漸近的にベイズ識別に近づく．パターン数が比較的少ない場合でも良い結果を与えるが，入力とすべてのパターンとの距離を計算する必要がある．単純パターンマッチング法であるユークリッド距離や相関係数は k-NN 識別法で $k=1$ とする場合になる．複合類似度法，部分空間法[13]は，単純パターンマッチング法がカテゴリの内部表現を特徴空間の点の分布としてとらえているのを，その空間の部分空間としてとらえる．識別は，入力の部分空間への射影を計算する方法であり，単純パターンマッチング法を拡張したものである．また，複合類似度を確率化するモデルが考えられている[14),15]．

連続音声中の音素はその置かれた音韻環境によってさまざまな変形を受け一つの音韻表現が多くの音形表現をもつ．たとえば鼻音化(金魚：/kingjo/, [kiŋjo])，無声化（来た：/kita/, [kita]）などがある．連続音声認識では，このような音韻表現から音形表現への変形を規則（音形規則：Phonological Rule)[16]として整理することが必要である．規則の利用法は，1)トップダウン的に単語辞書の基本形にあらかじめ適用しておく，2)ボトムアップ的にセグメンテーション，音韻認識をする段階で規則を適用して音韻系列を修正する，の二つの方式がある．変形は必ず起こるものや個人によって起こったり起こらなかったりするものもある．

(3) 単語認識

単語を表現する方法としては正書法の表現によるもの，音形規則を適用しネットワーク表現にしたものがある．また，単語辞書内で単語を単に並べるのでなく，効率よく検索できるように木構造にしたものがある．

これらの表現に従って単語の同定を行う際に，認識単位の認識尺度をもとに単語としてのスコアを計算する必要がある．その方法として木探索法または動的計画法 (DP: dynamic programming) が用いられる．動的計画法は可能なすべての組み合わせの計算を効率よく行う手法である[17]．時間軸を非線形に伸縮する時間正規化 (DTW: dynamic time warping) を行いながら照合する方式として，動的計画法を用いて照合することをＤＰマッチングと呼ぶ[18]．さらに，入力と各標準パターンをフレームごとに連続的にＤＰマッチングを行い，あるカテゴリの距離がしきい値以下になったときにそのカテゴリがマッチングした区間に存在すると判定することを連続ＤＰマッチングと呼ぶ[19]．これは入力中から単語などの単位を検出するのに使われる(ワードスポッティング)．また，全数探索ではないため最適解を得られない場合もあるが，効率のよい木探索法[20]も利用される．

単語認識ではさらに，1)音韻の付加，挿入，脱落の

誤り，2)単語境界の不確定さ，3)単語継続時間の変動と短い機能語（日本語の助詞，英語の冠詞，前置詞）の扱い，4)単語スコアとしてなにを使うか，などが問題となる．

（4） 話者適応・正規化処理

音声波形に含まれる個人性を考慮すると，不特定話者の認識の場合に比べて認識の性能は大幅に良くなることが知られている．そのためには発声者の声を登録し個人ごとに入力音声のスペクトル変動などを正規化したり認識システム側を発声音に適応化する方法が考えられる．登録のための音声が短い発声時間で済めばまったく登録しない現在の不特定話者の音声認識に比べ性能的にも使いやすいものとなる．

正規化では音声生成モデルに基づいた有声音区間における声帯波スペクトルと声道長が考えられる．声帯波スペクトルは音声スペクトルの全体の傾斜として，声道長は音声スペクトルの周波数軸方向の伸縮として現れ，たとえば男性の標準パターンで女性の入力を識別する実験では，線形な周波数軸上で線形伸縮が有効であることが示されている[21]．

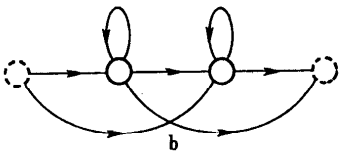
適応化では音素などの標準パターンを発声者に適応化して作成する方式の研究が行われている．教師なしで，あらかじめ用意した母音標準パターンの組から最適な組を選択し，それを基本として話者に適応化するもの[22]，ベクトル量子化（Vector Quantization：VQ）を利用したシステムの適応化[23]や複合類似度法を用いた認識システムにおける不特定話者辞書を基本とした適応化[24]などがある．

2.2 音声認識の最近の研究

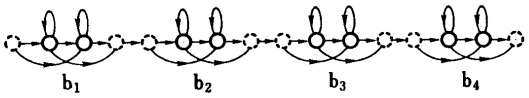
ここでは最近の研究動向として欧米の Hidden Markov Model の研究と日本の音節・音韻を単位とする研究に分けて概観する．

（1） 欧米の研究

欧米の連続音声認識の研究ではマルコフモデルを用いたものが多い．普通は入力音声を VQ などにより記号（符号）列に変換し，その記号列を発生するモデルにマルコフモデルを用いている．モデルのパラメータは記号発生確率と状態遷移確率である．マルコフモデルの例を図-1 に示す．このようなマルコフモデルでは記号列では内部状態の遷移が陽には見えないので Hidden Markov Model (HMM)[25] と呼ばれている．入力記号列の尤度は，入力記号列を発生する可能性のあるすべての状態遷移系列の起こる確率の和で表されるが，それを漸化的に計算するアルゴリズムとして Forward-Backward アルゴリズムが知られている．また，パラメータ推定法として，Baum-Welch の推定法により繰り返しながら収束状態を求め，そのときの値を結果とする．絶対的な意味での最適解に収束する保証はないのでパラメータの初期値の与え方が重要である．また，“焼きなまし”（Annealing）という考え方でよりよい局所最適解に収束させることも行われている[26]．さらに計算を簡単化して状態遷移の最大確率だけをもとめるのに Viterbi のアルゴリズムが使用されるがこれはＤＰマッチングに似たアルゴリズムになる[27]．マルコフモデルの各状態の継続時間のモデルを組み込む必要が明らかになり，いくつかのモデルの比較が行われている[28]．



（a） 音韻記号ｂを表すマルコフモデルの例



（b） 音韻表記が b_1, b_2, b_3, b_4 である単語ｗを表すマルコフモデルの例

図-1 マルコフモデルの例
丸印は状態，実線は状態遷移を表し各遷移ごとに音響ラベルを出力する

IBM では大語彙単語入力によるオフィス文書の口述筆記をタスクとしている[29),30]．従来から一貫して音声認識過程を通信理論の復号化（確率モデル）の問題にモデル化し，パラメータを統計的に自動学習して抽出する．２万語の語彙で一般的事務文書100文を（1698 単語）を認識した結果は，７名平均で 94.6% の認識率であった．

BBN では BYBLOS システムにおいて調音結合を考慮した HMM 音韻モデルを用いている[31]．システムには特定話者モードと話者適応モードがある．タスクであつかう文の構文を正規文法で表現し，約350語の語彙のタスクについて話者適応モードで15秒の訓練用発声により 97% の単語認識率を得た．

AT & T では継続時間を考慮した HMM[32] を音韻モデルとしている．音韻間構成を状態遷移マトリックス（マルコフチェイン），音韻パラメータとしてガウス過程モデル，単語辞書検索に HMM 中の状態系列

の部分列をキーとした redundunt hash-addressing[33] により単語ラティスを作成する．文認識部では単語の尤度，始端時刻，終端時刻をパラメータとして best-first 探索により認識を行っている．

（2） 日本の研究

英語の音節構造は子音-母音-子音（CVC）が基本であり音節数は約10000種程度である．一方，日本語では子音-母音（CV）が基本であり100種程度である．また，音節は仮名文字と対応しており，その特徴をいかして音節を単位とした認識による文入力が検討されている．音韻を単位とした認識では，不特定話者，母音の話者適応を行うものなどが研究されている．認識処理を大きく分けると音節・音韻ごとのセグメンテーションを行うものと，セグメンテーションを行わずに音節・音韻の照合を行うものがある．音節を単位とした認識の代表的な処理の流れを図-2 に示す．

セグメンテーションする場合，音節単位の認識ではエネルギ，スコア，スペクトル変化などから音節ごとに行うもの[24),34),35)]，比較的安定な母音特徴からVCV 単位に行うもの[36),37)] がある．しかし，このセグメンテーションは，音節の境界位置を厳密に決めるのではなく音節照合の範囲を決めるためのものである．音節の照合は連続ＤＰマッチングで行うもの，連続的に固定長で行うものがある．ＤＰマッチングを使うものでは，ＣＶの単位では非線形な伸縮をする程度は小さくする必要がある．連続的に固定長で照合する方式[24)]では音節の調音結合による変化，時間的変動などは複合類似度法の標準パターンで表現される．処理の例を図-3 に示す．音節認識部の結果は音節ラティスの形で出力され，言語処理部で形態素解析が行われる．また，音韻単位の認識では有声/無声などを判別する統計量に基づいたフィルタ出力やパワーを用いてセグメンテーションし，各区間の音韻群を決めた後に詳細な音韻識別する方法が考えられている[38)]．セグメンテーションを行う方式では言語処理部の負荷が軽減され語彙数は音節単位認識の場合に数万語程度まで扱われているがセグメンテーションの性能が認識率に大きく影響する．

セグメンテーションをしない場合，音節単位の認識では入力フレームごとにそれまでの入力の中で最適な音節候補，単語候補，文節候補をＤＰマッチングにより求めておく認識方式が研究されている[39)]．単語辞書は音節からなる木構造で表現される自立語，付属語からなり，文節内構文情報はオートマトンで表現されて

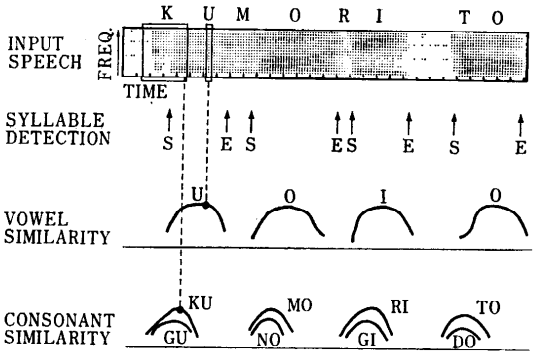


図-3 音声パターンと母音・子音類似度の例[24)]

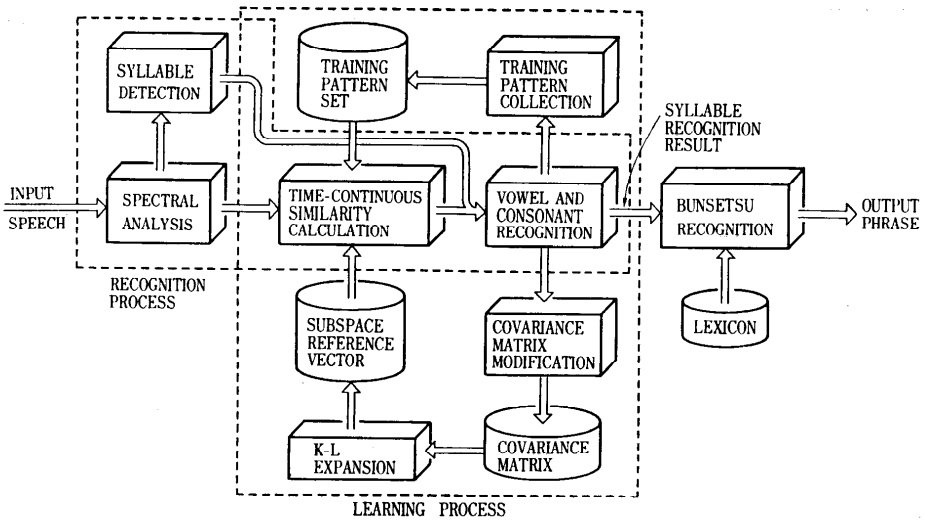


図-2 日本語文入力のための話者適応型音声認識のブロック図[24)]

いる．音節の標準パターンには無声化したものも含まれ，調音結合による変化にも対処している．また，音節を連続的に照合する場合にスコアとして事後確率を求め，時間的に変化する事後確率系列から音節ラティスを作り認識することが試みられている[40]．音韻単位の認識では，音韻ごとに用意した検出用判別フィルタの出力を選択して音韻記号系列を生成することが試みられている[41]．選択する際は広い時間窓の範囲を観測することにより音韻環境などを取り込んでいる．また，検出用判別フィルタの内部表現は区分的線形識別関数であり，判別フィルタと選択のための関数は繰り返し学習によって訓練する．セグメンテーションを行わない方式では処理量が多くなるため，一般には語彙数は限られる．

調音結合による音韻の変化に対処する方法として，従来から音響管モデルの共振周波数に対応するフォルマントをパラメータとして音韻環境の影響を補正する研究がある[42),43]．最近では音韻環境を考慮するトップダウン的音韻認識が試みられている[44]．これは入力語の仮説に，音韻の変形を考慮した音韻認識関数でスコアづけして認識する．また，調音モデルに基づく特徴パラメータなどにより前後の音韻環境の影響を補正して，対象音韻スコアづけをすることも試みられている[45]．さらに，変形を受けても特徴の相対的な位置関係が変わらないことを利用して弛緩法により認識すること[46]，音韻単位に比べてより細かい時間区分に分けた音形規則を用いること[47]，音形規則の抽出を対話的に行い利用する方法[48]なども検討されている．

3. 音 声 理 解

3.1 音声理解の技術項目

音声認識を人間から機械へのインタフェースの一つのチャネルと考えると，音声のメッセージ内容の理解が重要になる．しかも自然な会話では音声信号は話者の意図を表現するものとして生成される．この意図から文へ，文から音声へという段階には，常識の世界，場面の情報，文法，発声器官の物理的制約などの種々の規則・拘束がある．聞き手が人間の場合は，これらの情報を総合的に利用して内容を判断している．たとえば“l”と“r”の弁別ができない人でも英語をある程度理解できる．これは，必ずしも音韻レベルで完全に認識できなくても，場面の状況などにより発話内容を限定しているからである．機械の場合も，自然に発声された音声を理解するには，これらの情報や知識を総合的に取り入れることが自然である．人間は決められた文の単語列と全く同じでなくても，既知の言語ならその内容を理解することができる．このような理解機能をモデル化して実現しようとするのが音声理解システムといえる．音声理解システムには，**表-2** に示すような要素技術（知識源）の研究が含まれる．

（1） 音韻・単語認識

音韻認識・単語認識のレベルは連続音声認識の一部でもあり，根本的な違いはない．知識源として整理すると，音韻論レベルの知見として**2.**で述べた音形規則のほかに，単語が繋がったときに変形するものについては単語結合規則が考えられている．たとえば数字例“69”を繋げて発声すると「ロッキュー」となることがあるなどの現象を規則化するものである．

また単語認識のレベルで音声理解の重要な技術として，連続音声中の単語の位置を検出するワードスポッティング技術がある[19),49]．この音韻・単語レベルでのデータ構造として，位置と内容の双方に関して曖昧さを許すと音韻ラティスや単語ラティスと呼ばれる形式になる．

（2） 構文・意味・談話処理

要素技術としては，キーボード入力などの記号を直接処理対象とする自然言語処理[50]と基本的な違いはないが，音声理解にとって重要なものとして，話し言葉の文法の体系化や断片的入力の取り扱いがある．

構文レベルでは，文の解析技術が必要であり，句構造文法の各種の解析法が利用されることが多い．最近では単語や品詞の n 字組（n−gram）による候補単語

表-2 音声理解の技術項目

技術項目	内容	主要な課題
音韻・単語認識	連続音声の基本（表-1 参照）	ワードスポッティング技術
構文・意味・談話処理	係り受け解析などの文解析，省略の処理など	話し言葉の文法定式化，意味情報の表現と利用，省略の処理
韻律処理	基本周波数輪郭パターンなどの抽出と利用	文の構造（句境界や疑問文など）との関連性明確化
会話モデル	円滑な会話を行うためのモデル	利用者のモデル，発話の自然な確認方法
システム構成法	各処理部（知識源）の組み合わせ方と制御	合理的なスコアづけの方法

の絞り込みが行われている．これは文構造を n 個の単語の系列の出現確率で近似したものであり，単語同定の評価として確率を用いている場合には組み込みやすいという利点がある．また文節内構文規則の表現として確率オートマトンや3字組モデルを組み込んだものなどが研究されている．日本語文節の付属語部分は，一般に短い単語から構成されるために認識誤りが多い．そのため，構文駆動的な制御を行い，文節を一つの単位として取り扱うことも検討されている[51]．また述部を除いて文節の出現順序は比較的自由であり，文節相互の依存共起関係の解析方法や格形式・意味素性による共起関係について研究されている．

会話音声を理解する上で重要である話し言葉の文法は，会話固有の現象の整理が行われている段階である[52),53)]．

意味レベルでは，意味素性により単語の意味を表したり，意味ネットワークなどを利用して単語間の意味的な関連性を表現する方法が用いられる．大規模な意味的な辞書の整備が必要となっている．

談話レベルでは，省略や代名詞の照応参照の問題，発話の意図を抽出する発話行為の問題などを扱う．特に音声の対話では断片的な入力などの省略が多く，事前知識や会話の経緯から発話の意図を推論することが必要になる．

（3） 韻律処理

韻律情報とは，音声の基本周波数・パワー・音韻の継続時間長をさし，書き言葉での句読点に対応するとされる．これらの情報はアクセントやイントネーションを表す特徴となり，音声の自然性と深い関わりがある．音声認識・理解では句境界の検出などの手掛かりを与えてくれる[54)~56)]．たとえば，「ニワニワトリガイル」では，（分節的な）音韻内容だけでは「庭には鳥がいる」と「二羽鶏がいる」の区別は難しいが，基本周波数の時間軸上の谷が句境界に対応して出現するため，区別することができる．また疑問文の区別への手掛かりを与えてくれる可能性をもっている．

（4） 会話モデル

音声理解では最終的にユーザと対話できることを目標としているといえる．入力音声の認識判定結果に曖昧さが残る場合には，結果の確認のための応答が要求される[57)]．発話の確認にキーワードを使うことにより自然なやりとりで，しかも少ない回数で確認できることが報告されている[58)]．また会話を行うためには，入力に対して適切な応答をする必要がある．そのときには会話の原則（会話モデル）が考えられ，それに従った応答を行う必要がある[59)]．

（5） システム構成法

上記の各種知識源を組み合わせてシステムを構成する方法は，人間の音声知覚過程のモデル化そのものであり，音声理解システム全体の成否に繋がる重要な項目である．システム構成法として各種の知識源の組み合わせ方が考えられている[60)]．黒板モデルは，各レベルの知識源を独立した処理モジュールとして準備し，共通のデータベース（黒板）を通じ，知識源が通信しながら処理を進めていく．これは HEARSAY-Ⅱ システム[61)]で開発された方法で，各知識源が独立に動作できるため，分散処理に向いている．現在では，一般問題解決システムのモデルとして利用されている．

一般に各知識源の知見には不完全なところが残されており，それぞれの処理には曖昧さが残る．したがって，それらをどのように組み合わせて候補を評価するかというスコアづけの方法が重要になる．多種類の知識や処理モジュールを使うときのスコアづけの一つの方法として，事後確率を用いる方法が検討された[62)]．

また曖昧さが組み合わさるので，検討すべき候補数が膨大となり，動的計画法に代表される全数探索を行うことは不可能になる．したがって，最終的に最もスコアの大きいものを見つけるために，どのような順序で中間的な候補を調べていくかの戦略が重要になる．複数個の有望な候補を並列的に調べていくビーム探索法（beam-search）[63),64)]がよく利用される．

3.2 音声理解の最近の研究

最近の音声理解研究は，米国国防総省の戦略計算（Strategic Computing）プロジェクト[65)]，日本の ATR の自動翻訳電話プロジェクト[66)]，英国の ALVEY 計画などの国家的規模のプロジェクトとして実行されているほかに，各国の研究機関で活発に研究されている．最近の動向として，知識工学手法の利用と従来からの分散型モデルに分けられる．

（1） 知識工学的手法による音声理解

知識工学的な手法の代表的なものとしてプロダクションシステムを利用する方法[67)~69)]についてみる．これは if-then 型のルールを用意しておき，if 部の条件が一致するものがあれば，then 部の記号に書き替えていくものである．たとえば，

もし（if）破裂音が無声で，
かつ（and）破裂音の調音位置が歯茎である，
ならば（then）その破裂音は/ t /である．

といったルールを用意しておく．これらのルールは，人間がスペクトルの時系列データ（スペクトログラム）を目視して音韻記号を割り付けるときに利用している知識を明確化することにより作成する．50名の発声した連続音声中の/t/の部分の識別実験では，84～88%の精度が得られており，人間の視察判定結果と同水準であることが報告されている[68]．現在は連続音声認識として研究されているが，構文や意味の高次言語情報の処理を取り入れる上でも拡張性は高いといえる．

この方法の根拠としては，人間が言語情報を併用して目視によって音韻記号を割り付けると90%以上の高精度になることが知られている[70),71)]．しかし，どのようにして人間が用いる知識をプロダクションルールに表現し，濃淡画像でもあるスペクトログラムを扱うか，特に特徴の相対的時間関係の取り扱いや曖昧さに対するスコアづけの方法が研究課題である．類似のアプローチとして構文的認識法があるが，音声パターンの曖昧さに対処するためにファジィの考え方が導入された[72]．プロダクションシステムでは，MYCIN流の確信度（Certainty Factor）が用いられることが多いが，その根拠は必ずしも明確ではなく，今後の課題である．

また知識工学的な手法としてはほかにフレームを利用するもの[73]も考えられている．

（2） 分散モデルによる音声理解研究

3.1 で述べた各知識源の階層を分けた分散型あるいは階層モデルによるシステムの研究が続けられている．システム全体の構成では黒板モデルに代表されるような分散システムの枠組みを用意し，その中に各レベルの知識源を取り入れる形式のものが多い[74)~77)]．典型的なシステム構成例を**図-4**に示す．各レベルの知識源を独立に構成し，それらの間にはラティス形式の曖昧さを許したデータ構造をもつことにより処理を進めるものである．

一例として CMU で研究されている音声理解システム[78]をみる．システム構成は**図-5**に示されるように Hearsay II システムから受け継いだもので，知識源である処理部とそれらが並列に動作しながら参照できる黒板からなる．処理部には音韻識別部，単語仮説部，文仮説部などがあり，音韻識別部は，音韻クラスの位置候補を見付ける位置検出部と詳細な音韻識別を行う識別部からなり，音韻ラティスの形式で結果を黒板へ出力する．単語仮説部は，単語辞書と音韻ラティスの照合部を基本として，音響特徴のはっきりした音節核を検出し単語の照合範囲を狭める安定特徴抽出部，入力を無音区間や母音性区間に荒く分けるラベリング部，音韻ラティスの修正を行う音韻ラティス統合部，パーザが依頼する単語の検証を行う検証部から構

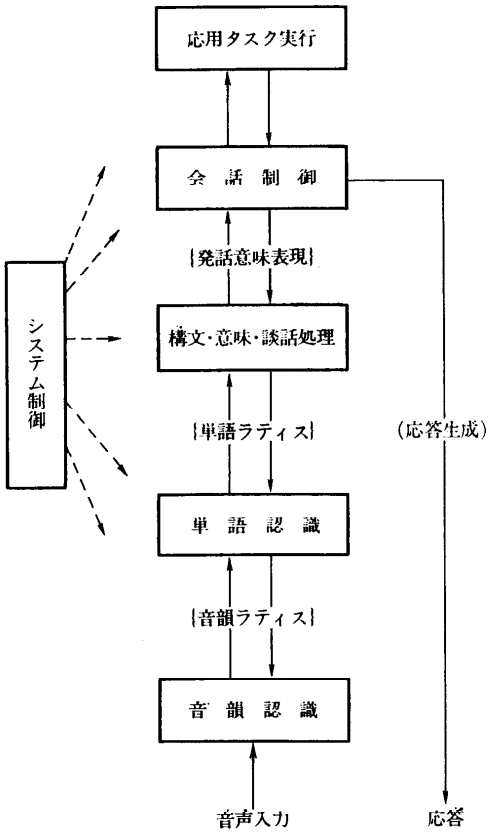


図-4 分散型音声理解システムの構成例

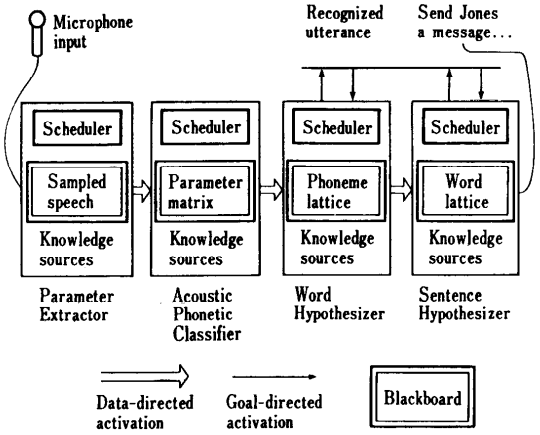


図-5 CMU の音声理解システムの構成[78]

成される．文仮説部は構文情報と意味情報を合わせて処理できるよう格フレーム分析により，単語の候補を解析する．このような構成により，ボトムアップ的に動作するようなシステムを開発している途中である．

分散モデルによる音声理解全般について，最近では構文レベルの処理として，単語ラティスや文節ラティスを効率的に解析する方法が活発に検討されている[79)~84)]．また対話のトピック管理までのレベルを一様な意味ネットワークに記述する方法[85)]，話題の取り扱い方法[86)]，会話制御の方法[87)]なども研究されている．**2.** で述べた HMM と知識工学的手法を黒板モデルで組み合わせる方式も提案されている[88)]．

4. 音声認識・理解研究の課題

2.3 で概観した音声認識・音声理解の研究には，表-1，表-2で示した各技術項目においてそれぞれの問題が存在する．

音韻認識においては，欧米で HMM が広く用いられている．日本の場合はＣＶ単位の認識が多い．しかし前後の環境の影響が無視できない場合も多く，パターンの変動がうまく表現できる HMM の利用がさらに活発に検討されるであろう．その際に音韻やＣＶ音節のモデルをおのおの複数個準備することが予想されるが，モデルのパラメータ推定に多くのデータが必要になる．現在データベースの整備が行われている[89)]が，連続音声の各種の音韻環境について整理されたデータの充実が望まれる．

最近は，話者の正規化に関する研究発表が比較的少なくなっているが，やはり音声認識の大きな課題であることに変わりはない．第一次のモデルとして周波数軸上のスペクトル変動が取り扱われたが，さらに時間軸上に現れる個人性特徴の明確化やその正規化が重要である．

日本語の構文レベルの課題として，付属語の取り扱いがある．発声が相対的に不明瞭になる文節の末部にあり，しかも助詞・助動詞の各単語は短いため，認識が難しい．付属語の部分は単にトップダウン的に検証するのに止めるなど，自立語とは異なった取り扱いを検討する必要がある．

音声理解システムとしては，低次の連続信号を処理するレベルでは数学的な手法を用い，また高次の言語情報の利用には，AI 的な手法が用いられる場合が多い．これらをいかに組み合わせるか，あるいは両方の性格を扱える統一的な手法の開発が望まれる．この点，知識工学的手法は，全知識源を統一的に扱える可能性をもっているが，信号レベルの処理方法と曖昧さの扱いが十分検討される必要がある．システムを構成する際にはボトムアップ的に得られる距離・類似度とトップダウン的に指定される確信度との組み合わせを評価するスコアづけが重要な研究課題である．

音声理解の研究項目の中には，自然言語理解の研究と重なるものも多い．自然言語理解が可能にならないかぎり，音声理解は不可能であるともいえるが，逆に膨大な曖昧さの合理的な扱いの中にこそ，有用な知覚モデルを構築できる可能性がある．音声理解の研究としては，このような曖昧さを扱うシステマティックな手法の開発，通信メディアとして音声の優位性を主張できる会話の問題，音声固有の情報である韻律情報の扱いなどが，連続音声認識技術の研究とともにとくに重要であろう．

また本稿では述べなかったが，音声認識技術の実用化を妨げる大きな壁として，環境雑音の対策が残されている．これには音声区間の検出，認識対象語と不要音の区別などの問題が含まれている．

5. むすび

音声認識及び理解研究の現状を概観し，今後解決すべき課題を整理した．音声認識としては連続単語認識の研究が一段落し，現在は音韻認識が活発に研究されている．ここでは，調音結合の正規化，話者の正規化を初めとした研究課題が山積している．また音声理解としては，個別の要素技術である各種知識源の深耕が行われている．知識源が不十分であるがゆえに，その曖昧さの統一的な取扱いが重要になる．現在，聴覚モデルの本格的な検討が開始されている[90)]が，認識と理解を通じてその成果に期待されるものも多い．

音声は人間のもっとも基本的な通信媒体であるとともに，実用レベルでは安価なキーボードやマウスを初めとした他の媒体と競合するところも多い．今後の研究により，これらの課題が一日も早く解決され，人間にとってもっとも自然なメディアにより機械と対話できる日がくることを望みたい．

参考文献

1) 例えば，三浦種敏（監修）：新版聴覚と音声，電子通信学会編，コロナ社（1980）．
2) Makhol, J.: Linear Prediction: A Tutorial Review, Proc. IEEE, Vol. 63, No. 4 (1975).
3) Markel, J. D. and Gray Jr., A. H.: Linear Prediction of Speech, Springer-Verlag (1976).

4) 板倉, 斎藤：偏自己相関係数による音声分析合成系, 音講論 2-2-6 (1969).
5) Itakura, F., Saito, S., Koike. Y., Sawabe, H. and Nishikawa, M.: An Audio Response Unit based on Partial Correlation, IEEE Trans., Vol. COM-20 (1972).
6) 例えば, Reddy, D.R. and Vicens, P.J.: A Procedure for the Segmentation of Connected Speech, J. Audio Eng. Soc., Vol. 16, pp. 404-412 (1968).
7) Bridle, J.S. and Sedgwick, R.: A Method for Segmentoing Acoustic Patterns with Applications to Automatic Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 656-659 (1977).
8) 例えば, Dixon, N.R. and Silverman, H.F.: The 1976 Modular Acoustic Processor (MAP), IEEE Trans., Vol. ASSP-25, No. 5, pp. 367-379 (1977).
9) 杉山, 鹿野：ピークに重みをおいた LPC スペクトルマッチング尺度, 信学論, Vol. J65-A, No. 9, pp. 965-972 (1982).
10) 飯島泰蔵：パターン認識, 電気工学体系 43, コロナ社 (1973).
11) Itakura, F. and Umezaki, T.: Distance Measure for Speech Recognition based on the Smoothed Group Delay Spectrum, Proc. Int. Conf. Acoust. Speech, Signal Process., 29. 1 (1987).
12) Cover, T.M. and Hart, P.E.: Nearest Neighbor Pattern Classification, IEEE Trans., Inform. Theory, Vol. IT-13, pp. 21-27 (1967).
13) Watanabe, S. and Pakvasa, N.: Subspace Method in Pattern Recognition, Proc. 1st Int. Conf. Pattern Recog., pp. 25-32 (1973).
14) 瀬川英生：複合類似度法における類似度値の分布について, 信学技報, PRU 87-18 (1987).
15) 浮田, 新田, 渡辺：統計的単語同定法を用いた不特定話者連続音声認識, 信学論, Vol. 68-D, No. 3, pp. 284-291 (1985).
16) Oshika, B.T., Zue, V.W., Neu, H. and Aurbach, J.: The Role of Phonological Rules in Speech Understanding Research, IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-23, No. 1, pp. 104-112 (1975).
17) Bellman, R.: Dynamic Programming, Princeton Univ. Press (1957).
18) 迫江, 千葉：動的計画法を利用した音声の時間正規化に基づく連続単語認識, 音響誌, Vol. 27, No. 9, pp. 483-490 (1971).
19) Christiansen, R.W. and Rushforth, C.K.: Detecting and Locating Key Words in Continuous Speech Using Linear Predictive Coding, IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-25, No. 5, pp. 361-367 (1977).
20) Nilsson, N.J.: Problem-Solving Methods in Artificial Intelligence, McGraw-Hill (1971).
21) 松本, 脇田：Frequency Warping による話者正規化, 音声研資, S79-24 (1979).
22) 杉山, 好田：認識結果を利用した母音標準パターンの教師なし話者適応法, 信学論, Vol. J69-D, No. 8, pp. 1197-1204 (1986).
23) Shikano, K., Lee, K.F. and Reddy, R.: Speaker Adaptation through Vector Quantization, Proc. Int. Conf. Acoust., Speech, Signal Process., 49.5 (1986).
24) Tsuboi, H., Takebayashi, Y., Matsu'ura, H., Nitta, T. and Hirai, S.: Speaker-Adaptive Connected Syllable Recognition Based on the Multiple Similarity Method, Proc. Int. Conf. Acoust., Speech, Signal Process., 49.8 (1986).
25) Rabiner, L.R. and Jung, B.H.: An Introduction to Hidden Markov Models, IEEE ASSP Magazine, 3, 1, pp. 4-16 (1986).
26) Pual, D.B.: Training to HMM Recognition by Simulated Annealing, Proc. Int. Conf. Acoust., Speech, Signal Process, 1.4 (1985).
27) Juang, B.H.: On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition-A Unified View, AT & T Bell Lab. Tech. Journal, Vol. 63, No. 7, pp. 1213-1243 (1984).
28) Russell, M.J. and Cook, A.E.: Experimental Evaluation of Duration Modelling Techniques for Automatic Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., 20.19 (1987).
29) Bahl, L.R., Jelinek, F. and Mercer, R.L.: A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-5, No. 2, pp. 179-190 (1983).
30) Averbuch, A., Bahl, L., Bakis, R., Brown, P., Daggett, G., Das, S., Davis, K., De Gennaro, S., De Souza, P., Epstein, E., Fraleigh, D., Jelinek, F., Lewis, B., Mercer, R., Moorhead, J., Nadas, A., Nahamoo, D., Picheny, M., Shichman, G., Spinelli P., Van Compernolle, D. and Wilkens, H.: Experiments with the Tangora 20000 Word Speech Recognizer, Proc. Int. Conf. Acoust., Speech, Signal Process., 17.3 (1987).
31) Chow, Y.L., Dunham, M.O., Kimball, O.A., Krasner, M.A., Kubala, G.F., Makhoul, J., Price, P.J., Roucos, S. and Schwartz R.M.: BYBLOS: The BBN Continuous Speech Recognition System, Proc. Int. Conf. Acoust., Speech, Signal Process., 3.7 (1987).
32) Levinson, S.: Continuous Speech Recognition by means of Acoustic/Phonetic Classification Obtained from a Hidden Markov Model, Proc.

Int. Conf. Acoust., Speech, Signal Process., 3.8 (1987).
33) Kohonen, T., Riittinen, H., Jalanko, M. and Haltsonen, S.: A Thousand-Word Recognition System Based on the Learning Subspace Method and Redundant Hash Addressing, Proc. 5-th IJCPR, pp. 158-165 (1980).
34) Sugamura, N., Tsuboi, T. and Nakatsu, R.: Japanese Text Input System Based on Continuous Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., 21.16 (1987).
35) Togawa, F., Hakaridani, M., Iwahashi, H. and Ueda, T.: Voice-Activated Word Processor with Automatic Learning for Dynamic Optimization of Syllable-Templates, Proc. Int. Conf. Acoust., Speech, Signal Process., 21.25 (1986).
36) Watanabe, T.: Syllable Recognition for Continuous Japanese Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., 42.18 (1986).
37) Hataoka, N., Amano, A. and Yajima, S.: VCV Segmentation and Phoneme Recognition in Continuous Speech, Proc. Int. Conf. Acoust., Speech, Signal Process., 42.19 (1986).
38) Makino, S., Homma, S. and Kido, K.: Speaker Independent Word Recognition System Based on Phoneme Recognition for a Large Size (212 word) Vocabulary, J. Acoust. Soc. Japan (E), Vol. 6, No. 3, pp. 171-180 (1985).
39) 伊藤，中川：音節パターンを用いたオートマトン制御拡張連続ＤＰ法による文音声の認識，信学技報，SP 86-62 (1986).
40) 藤崎，広瀬，井上，山下：音節標準パターンを用いた音声認識の一方式，音声研資，S 83-16 (1983).
41) Moriai, S., Makino, S. and Kido, K.: Phoneme Recognition in Continuous Speech using Phoneme Discriminant Filters. Int. Conf. Acoust., Speech, Signal Process., 42.7 (1986).
42) 桑原，境：連続音声中の母音連鎖における調音結合効果の正規化，音響誌，Vol. 29, No. 2, pp. 91-99 (1973).
43) 佐藤，藤崎：ホルマント周波数上での調音結合の定式化と音声自動認識への適用，音響誌，Vol. 34, No. 3, pp. 177-185 (1978).
44) 松永，鹿野：Top-Down 音韻認識と Bottom-Up 音韻認識を融合した音声認識，信学論，Vol. 68-D, No. 9, pp. 1641-1648 (1985).
45) Shirai, K., Kobayashi, T. and Yazawa, J.: Estimation of Articulatory Parameters by Table Look-up Method and Its Application for Speaker Independent Phoneme Recognition, Int. Conf. Acoust., Speech, Signal Process., 42.6 (1986).
46) Yu, R. and Kimura, M.: A Relaxation Technique for Seeking Optimal Vowel Candidate Sequence, Proc. Int. Conf. Acoust., Speech, Signal Process., 49.13 (1987).
47) 速水，田中，太田：規則と VCV・CVC 環境別標準パターンによる音韻の認識変動記述と単語認識実験，信学技報，SP 86-76 (1986).
48) Kimura, S. and Nara, Y.: Extraction of Phonemic Variation Rules in Continuous Speech Spoken by Multiple Speakers, Int. Conf. Acoust., Speech, Signal Process., 20.8 (1987) st., Speech, Signal Process., 49.13 (1987).
49) 川端，好田：音声の定常部と過渡部における継続時間長伸縮の違いを考慮したワードスポッティング，信学論，Vol. J 69-A, No. 2, pp. 261-270 (1986).
50) 入門書として，テナント，H，森　健一他訳：自然言語処理入門，産業図書（昭 59），長尾　真：言語工学，昭晃堂（昭 58），などがある.
51) 岡田，伊藤，松尾，牧野，城戸：日本語 DICTATION SYSTEM における日本語処理，信学技報，SP 86-33 (1986).
52) 国立国語研究所報告 18：話し言葉の文型（1）—対話資料による研究—(1960)，国立国語研究所報告 23：話し言葉の文型（2）—独話資料による研究—(1963).
53) 有田，小暮，野垣内，前田，飯田：メディアに依存する会話の様式，情報処理学会研究会資料，87-NL-61-5 (1987).
54) Lea, W. A., Medress, M. F. and Skinner, T. E.: A Prosodically Guided Speech Understanding Strategy, IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-23, No. 1, pp. 30-38 (1975).
55) 浮田，中川，坂井：日本語算術文の音声認識におけるピッチパターンの利用，信学論，Vol. 63-D, No. 11, pp. 954-961 (1980).
56) Waibel, A.: Prosodic Knowledge Sources for Hypothesization in a Continuous Speech Recognition System, Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 856-859 (1987).
57) 好田，中津，鹿野，伊藤：音声によるオンライン質問回答システム，音響誌，Vol. 34, No. 3, pp. 194-203 (1978).
58) 浮田，石川，中川，坂井：音声による対話システムにおける発話の確認方法，情報処理学会論文誌，Vol. 22, No. 6, pp. 589-595 (1981).
59) Hayes, P. and Reddy, D. R.: An Anatomy of Graceful Interaction in Spoken and Written Man-Machine Communication, Carnegie-Mellon Univ. Rep. CMU-CS-79-144 (1979).
60) Reddy, D. R. and Erman, L. D.: Tutorial on System Organization for Speech Understanding, in D. R. Reddy (Ed.) Speech Recognition, Invited Papers Presented at the 1974 IEEE Symposium, Academic Press (1975),

pp. 457-479.
61) Erman, L. D., Hayes-Roth, F., Lesser, V. R. and Reddy, D. R.: The Hearsay II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty, ACM Computing Surveys, Vol. 12, No. 2, pp. 213-254 (1980).
62) Woods, W. A.: Optimal Search Strategies for Speech Understanding Control, Artificial Intelligence, Vol. 18, pp. 295-326 (1982).
63) Lowerre, B. and Reddy, D. R.: The Harpy Speech Understanding System, in W. A. Lea (Ed.) "Trends in Speech Recognition", pp. 340-360, Prentice-Hall (1980).
64) Nakagawa, S. and Sakai T.: A Parallel Tree Search Method, Proc. 6-th IJCAI, pp. 628-632 (1979).
65) Stefik, M.: Strategic Computing at DARPA: Overview and Assessment, Comm. ACM., Vol. 28, No. 7, pp. 690-704 (1985).
66) 鹿野, 樽松: 音声理解研究の動向, 文献 99) 中, pp. 948-952.
67) Newell, A.: Harpy, Production Systems, and Human Cognition, in Cole, R. A. (Ed.) Perception and Production of Fluent Speech, Lawrence Erlbaum Associates, Publishers, New Jersey (1980).
68) Zue, V. W. and Lamel, L. F.: An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition, Proc. ICASSP '86, 23. 2 (1986).
69) 溝口, 田中, 福田, 辻野, 角所: 連続音声認識エキスパートシステム-SPREX-, 信学論, Vol. J 70-D, No. 6, pp. 1189-1198 (1987).
70) Klatt, D. H. and Stevens, K. N.: On the Automatic Recognition of Continuous Speech: Implication from a Spectrogram-Reading Experiment, IEEE Trans. Audio and Electroacoustics, Vol. AU-21, No. 3, pp. 210-217 (1973).
71) Zue, V. W. and Cole, R. A.: Experiments on Spectrogram Reading, Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 116-119 (1979).
72) De Mori, R. and Laface, P.: Use of Fuzzy Algorithms for Phonetic and Phonemic Labeling of Continuous Speech, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 2, pp. 136-148 (1980).
73) Carbonell, N., Damestoy, J., Fohr, D., Haton, J. and Lonchamp, F: APHODEX, Design and Implementation of an Acoustic-Phonetic Decoding Expert System, Proc. Int. Conf. Acoust., Speech, Signal Process., 23. 3 (1986).
74) Alleva, F., Bisiani, R., Forin, S. and Lerner, R.: A Distributed System Architecture for Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., 30. 4 (1986).
75) Martelli, T., Miclet, L. and Tubach, J.: REMORA A Software Architecture for the Collaboration of Different Knowledge Sources in Phonetic Decoding of Continuous Speech, Proc. Int. Conf. Acoust., Speech, Signal Process., 10. 8 (1987).
76) Kobayashi, Y. and Niimi, Y.: A New Architecture of a Speech Understanding System-A Hybrid of a Hierarchical and a Network Models, Proc. Int. Conf. Acoust., Speech, Signal Process., 30. 7 (1986).
77) Shigenaga, M., Sekiguchi, Y., Yagisawa, T. and Kato, K.: A Speech Recognition System of Continuously Spoken Japanese Sentences and an Application to a Speech Input Device, Proc. Int. Conf. Acoust., Speech, Signal Process., 30. 5 (1986).
78) Adams, D. A. and Bisiani, R.: The Carnegie-Mellon University Distributed Speech Recognition System, Speech Technology, MAR/APR., pp. 14-23 (1986).
79) Tomita, M.: An Efficient Word Lattice Parsing Algorithm for Continuous Speech Recognition, Proc. Int. Conf. Acoust., Speech, Signal Process., 30. 3 (1986).
80) 好田正紀: 文節ラチス上で係り受けの整合性を考慮したオートマトン制御の文節列選択アルゴリズム, 信学論, Vol. J 70-D, No. 8, pp. 1571-1578 (1987).
81) Ney, H.: Dynamic Programming Speech Recognition Using a Context-Free Grammar, Proc. Int. Conf. Acoust., Speech, Signal Process., 3. 2 (1987).
82) Stern, R. M., Ward, W. H., Hauptmann, A. G. and Leon J.: Sentence Parsing with Weak Grammatical Contraints, Proc. Int. Conf. Acoust., Speech, Signal Process., 10. 6 (1987).
83) Nakagawa, S.: Spoken Sentence Recognition by Time-Synchronous Parsing Algorithm of Context-Free Grammar, Proc. Int. Conf. Acoust., Speech, Signal Process., 20. 9 (1987).
84) 新美, 小林, 渦原: 音声理解システムにおける言語情報のトップダウン的利用方式, 信学論, Vol. J 70-D, No. 9, pp. 1772-1782 (1987).
85) Niemann, H., Brietzmann, A., Ehrlich, U. and Sagerer, G.: Representation of a Continuous Speech Understanding and Dialog System in a Homogeneous Semantic Net Architecture, Proc. Int. Conf. Acoust., Speech, Signal Process., 30. 6 (1986).
86) Kobayashi, T. and Shirai, K.: A Network Model Dealing with Focus of Conversation for Speech Understanding System, Proc. Int.

Conf. Acoust., Speech, Signal Process., 30.8 (1986).
87) Alinat, P., Gallais, E., Haton, J., Pierrel, J. and Richard, P.: A Continuous Speech Dialog System for the Oral Control of a Sonar Console, Proc. Int. Conf. Acoust., Speech, Signal Process., 10.3 (1987).
88) Haton, J., Carbonell, N., Fohr, D., Mari, J. and Kriouille, A.: Interaction between Stochastic Modeling and Knowledge-Based Techniques in Aoustic-Phonetic Decoding of Speech, Proc. Int. Conf. Acoust., Speech, Signal Process., 20.20 (1987).
89) 板橋秀一：音声データベース，文献100)中，pp. 433-438.
90) 古井貞熙：音声知覚研究とその音声情報処理への応用，文献99)中，pp. 953-958.
入門書，特集記事として次のものがある．
91) 新美康永：音声認識，共立出版（昭54).
92) 中川聖一：音声情報の認識と理解，坂井利之編著，情報基礎学詳説，第4章，コロナ社（昭58).
93) 中田和男：音声，日本音響学会編，音響工学講座，コロナ社（昭52).
94) 斉藤収三，中田和男：音声情報処理の基礎，オーム社（1981).
95) 古井貞熙：ディジタル音声処理，東海大学出版会（1985).
96) 特集 Man-Machine Communication by Voice, Proc. IEEE, Vol. 64, No. 4 (1976).
97) 特集 Man-Machine Speech Communication, Proc. IEEE, Vol. 73, No. 11 (1985).
98) 特集 音声情報処理，情報処理，Vol. 24, No. 8 (1983).
99) 特集 音声情報処理の最近の動向，音響誌，Vol. 42, No. 12 (1986).
100) 音声処理特集，電子情報通信学会誌，Vol. 70, No. 4 (1987).

（昭和62年11月30日受付）