解　説

機械による会話音声の認識・理解研究の動向*

中　川　聖　一**

1. はじめに

機械による音声の自動認識の研究は，大きく分けて次の四つのレベルで行われている．

1. 音韻（音素）・音節の認識．
2. 単語音声の認識．
3. 会話音声の認識・理解．
4. 話者認識．

普通，音声認識といえば，言語情報の認知といった狭い意味で使われるので，話者情報の認知のことを特に話者認識といっている．

これらは，**図-1** のように互いに関連しており，会話音声の認識・理解は，音声認識のすべての問題を包括しているが，ここでは会話音声の認識・理解における特有の問題と話者認識についての最近の研究動向について述べる．

音声認識に言語情報を導入する試みは古くから進められていた．Denes や 堂下による音韻遷移確率の利用[1),2)]，Alter によるフォートラン語の音声認識における構文情報の利用[3)]，Reddy・Vicens による非常に制限された自然言語文の音声認識[4)] 等が先駆的な研究である．これらの研究を経て，Fant[5)] や Reddy[6)] によって言語情報を用いた自然言語文の音声認識システムのモデルが提案され，1971 年以降，米国を中心として音声理解システムの名のもとに各国で活発な研究が進められてきた[7)~14)]．このような活発な研究にもかかわらず，すぐ実用化できるような自然言語文の音声認識・理解システムは開発されず，改めて音声認識の問題の深さ・難しさが浮き彫りにされた．しかし，音声認識における言語情報の表現・蓄積・検索・制御等に関する人工知能面での研究が進み，他の分野，特に画像理解システム研究にその経験と成果が利用されている[15)]等多くの成果のあったことも事実である．

筆者の些細な経験と見聞により，これらの研究を振り返ることによって最近の研究方向を読者に伝え，同時に音声理解研究は，以前の機械翻訳研究と異なって小規模なタスクの設定が可能であり，実用に耐えうるシステムの開発が可能なことを述べたい．

なお，音声情報処理の応用分野として音声合成・音声認識と同様に重要である話者認識についても述べるが，紙数の都合で，具体的な認識方法の記述は避け，個人性を特徴づけるパラメータについての最近の研究成果の解説にとどめることを断っておく．

2. 音声理解システム（SUS）とは

図-1 からもわかるように，会話音声の認識は孤立単語音声の認識をベースにしている．会話音声には，無意味語（えー，あのー とかの間投詞や助詞の引き伸ばし）が含まれうるが，それを除けば連続単語音声とみなせ，一見連続単語音声の認識手法と全く同様の手法が適用できると考えられるかも知れない．しかし，

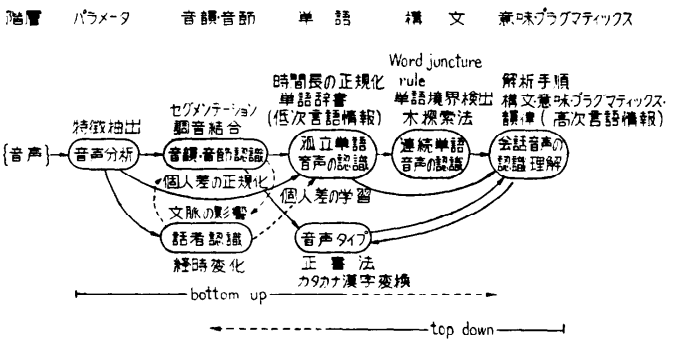

図-1 音声認識の階層と関連する諸問題

* Research Trends in Speech Recognition and Understanding by Machine by Sei-ichi NAKAGAWA (Department of Information Science, Faculty of Engineering, Kyoto University).

** 京都大学工学部情報工学科

会話音声の場合は，認識対象語彙数は，たとえタスク*を限定しても数百単語程度考えておく必要があるが，現在の研究レベルでは，100 単語の孤立単語音声ですら，不特定話者に対して 90% の認識率を得ることは容易ではない．今仮りに 90% の認識率を得るシステムがあるとしても，1文が 10 単語からなる会話音声なら認識率は 40% 以下(0.9^{10})になってしまう．これからもわかるように，会話音声を連続単語音声と同様な発想で扱う限り研究対象ですらなりえない．ところが幸いに，会話音声には冗長な情報が含まれているから，これらを援用すれば限定された世界なら十分扱える対象となってくる．これがいわゆる高次言語情報で，構文・意味・プラグマティックス・韻律情報などがある．これらを効果的に用いるためには，「自然言語の理解」という大きな困難な問題に取り組まなくてはならない．音声理解研究は自然言語理解研究と比べタスクが限定されているとはいえ，当然解析手順をどうするかとか相互の情報交換をどうするかとかの人工知能面での研究も必要となってくる．このような種々の知識源を統一的に扱いながら，かつ入力音声と言語符号を一対一に対応させることは非常に難しいので，音声理解という新しい概念が生まれた．つまり，入力が会話音声であることから発話文の正確な認識を必ずしも要求せずに，発話内容の理解で十分とする．この点が従来の音声認識との根本的な違いで，音声理解(Speech Understanding)と呼んで区別している[7]．

このような言語情報が，音声認識・理解にとっていかに重要であるかは，Klatt らのスペクトラムの読み取り実験が示してくれる[16]．彼らの実験によると，音声研究の専門家でも，音韻レベルの変換率は，10% が検出不能，17% が誤変換，40% が不完全な変換であった．それが，コンピュータの助けを借りながら，この不完全な変換系列から 200 語の辞書と構文・意味等の知識を用いて変換作業を繰り返すと 90% の単語が正しく同定された．

図-2 に，典型的な音声理解システムの構成図を示す．勿論個々のシステムによって多少異なった構成になっているが，ほとんどの場合はこの構成図に準じている．

音声理解システムが開発されれば，実用面での価値はいうまでもないが，学術面での価値として Newell は次の5つを挙げている[11]．

① 人間の音声認識過程の解明につながる．

② 調音結合・音韻論的現象・談話などの音声信号に与える種々の影響の形式化をもたらす．

③ 多種の知識源を持つシステムの実例を AI 研究に与える．

④ 「人間は容易に音声を認識できるけれども，機械には大変難しい」という神話をくずすことができる．

⑤ SUS に関する研究成果が，10 年前の電子技術やここ数年間のミニコンが果たしたと同様な役割（技術革新）を音声研究に与える．

音声理解システムを開発するに先だっては，システム設計に重要な影響を与える次の仕様を決めておく必要がある[7]．

① 連続音声を対象とするのか．
② 多数の話者を対象とするのか．
③ 発声者にどの程度の方言を許すのか．
④ 静かな部屋で発声するのか．
⑤ 性能のよいマイクロフォンを使うのか．
⑥ 発声者ごとに装置を調整してよいのか．
⑦ 発声者に装置への適用を求めるのか．
⑧ 語彙数はいくらか．
⑨ 構文は自然なものか．
⑩ タスクは何か．
⑪ 発話内容の理解率はいくらくらいにするのか．
⑫ 発話内容を理解する時間はどの程度か．
⑬ いつまでに開発するのか．

さらに，上述の各項目を決定したとしても，システムを具体的に構成するには，次のことを決める必要がある．

① 音響分析法（零交差，相関，スペクトラム，フォルマント，ケプストラム，LPC (Linear Predictive Coefficient, 線形予測係数) など）．

② 単語同定法（パターンマッチング，構造的，確率・統計的など）．

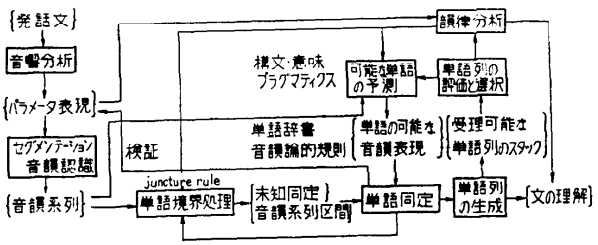

図-2 音声理解システムの典型的な構成例

* task：システムが入力として扱う対象．たとえば座席予約やニュースなどの情報検索

③ 木探索法 (depth-first, breadth-first, best-first, branch-and-bound など).

④ 文法の表現法 (変形文法, 文脈自由文法, 文脈依存文法, 遷移網文法, 依存文法など).

⑤ 知識源の表現法 (意味ネットワーク, 遷移ネットワーク, プロダクションシステム, 手続き埋め込み型など).

⑥ 解析手順 (左→右, 右→左, 両方向, top-down, bottom-up など).

⑦ 制御法 (階層モデル, 生成モデル, 集中型モデル, 分散型モデルなど).

これらを任意に選択することによって, 数千以上の SUS が設計可能であり, さらにタスクの規模によって最適なアプローチは異なるので, その優劣をすべて較べることはできないから, システム設計には, 多くの経験と事前調査並びにすぐれた直感力が要請されよう. 特に, 音響レベルの能力如何が, 言語レベルでの採るべき手法を決定づけることもあるので, まず音響レベルがどの程度の能力を持つかを見当づけてから設計にかかることが望ましい. いずれにしても, 開発段階では, 種々の手法が試行できるように柔軟なシステム構成を採用する必要がある.

3. ARPA の SUS プロジェクト

3.1 プロジェクトの経過

アメリカの ARPA (国防省の Advanced Research Projects Agency) の Roberts の要請により, 1970 年の春に音声理解システムの開発の可能性を討議するために, Newell を議長とする 9 人からなる研究グループが組織され, 以後 3 回の会合の後, 1970 年後半にその最終報告がまとめられた. この中で,「音声理解システムの開発は, 困難が予想されるが, 5 年間の研究期間があれば, ① 対象は連続音声で, ② 多数の話者の音声を, ③ 標準アメリカ英語を話す協力的な話者によって, ④ 静かな環境で発声され, ⑤ 品質のよいマイクロフォンを使用して, ⑥ 話者ごとにわずかの調整を許し, ⑦ かつユーザ自身に自然な適応を求める. ⑧ 使用する語彙は, ある程度選択された 1000 単語で, ⑨ 高度に人工的な構文構造を用い, ⑩ タスクはデータ管理や計算機の動作状態を問合せるような内容のものとし, ⑪ 意味上の誤りは 10% 以下におさえ, ⑫ 実時間の数倍で処理できるものを, ⑬ 1976 年にデモンストレーションできる」と報告した.

この報告を受けて, ARPA は同国内の大学・研究所と 1971 年秋から 5 カ年 (前期 2 年半, 後期 2 年半) 契約で資金援助 (毎年 300 万ドル) を行うことになった. 主契約先は, BBN (Bolt Beranek and Newman Inc.), CMU (Carnegie-Mellon University), Lincoln Lab. (MIT), SDC (System Development Corp.), SRI (Stanford Research Institute) の 5 研究機関, 補助契約先は, SCRL (Speech Communication Research Lab.), Speech Communication Lab., UNIVAC (at St. Paul) の 3 研究機関であった.

契約の前半終了時の 1973 年 11 月に各研究機関のデモンストレーションが行われ, 研究内容は 1974 年 4 月の CMU での音声認識シンポジウムで発表された[9)~11)]. この時点で, 契約先は比較的研究が進んでいた CMU, BBN および SRI と SDC との共同の 3 グループにプロジェクトは選択統合されて継続し, それぞれ 1976 年 8 月 31 日 (SDC), 9 月 8 日, (CMU), 9 月 10 日 (BBN) に最終デモンストレーションが行われた. その運営委員会は,「プロジェクトは, CMU の HARPY システムが初期の目標を達成して終了した.」と報告している[17)]が, 実験規模が小さくはなはだ疑問である.

同委員会は, プロジェクトで得られた知識や技術として次のものを挙げている. ① 各レベルにおける音声知識のネットワーク表現, ② 音韻論的規則, ③ 音響パラメータ表現としての LPC (線形予測係数) 技術, ④ タスクの困難度とか複雑度の測定技術, ⑤ 不確定なあるいは欠除単語を含む文の解析法, ⑥ 知能系の制御法, ⑦ 解析手順における優先度の決め方や探索順序の決定法.

次節以下では, 主に上述の 3 研究グループの 5 年間の研究経過を述べ, 研究の流れや方向をつかみ, 今後の音声理解研究の指針をさぐりたい. なお, 比較的たやすく入手できるまとまった文献として[10),11),13)], をあげておく (最近, 新しい報告が二・三でた[42),43)]).

3.2 CMU

このプロジェクトには, 最盛期で 18 名が参加していたそうで, 5 年間の研究経過は, SUS の開発→評価→拡張・改良→新 SUS の開発というサイクルが繰返され, いかにシステムが洗練されていき, 思想・アルゴリズムが収斂して行ったかがよくわかる.

HEARSAY-I は, Reddy らがスタンフォード大学で行っていた「積木の世界」における自然言語文の音声認識[4)]を基礎にしており, 特に音響分析では目新しい点は見当たらないが, 構文・意味・プラグマティクス

というような高次言語情報を積極的に導入し，音響情報も含めてそれが互いに独立に動作しうるBlackboardモデルや仮説・検定手法を採用した点等で高く評価され，以後の SUS 研究に指導的な役割を果した．ここで選ばれたタスクは「音声によってチェスを指す」もので，語彙数は31と少数であり，同じ CMU の Gillogly が開発したチェスプログラムを意味解析部に利用できたこと等，最初のタスクとしては的を得たものであり，かつ以後のシステムと同様，種々のタスクに適用できる汎用性のあるシステムであった．ただし，音響分析の不完全さや best first 法の採用，構文情報としては2字組や3字組程度の利用にとどまったこと，認識（理解）率が期待されたほど高くなかったこと（音響・構文・意味情報を使った場合で約80%，音響・構文情報だけを使った場合は約 50% の認識率．DESCAL と呼ばれる37語彙からなる算術文の認識では22%の認識率）等多くの問題が残った．

DRAGON システムは，音声理解システムにパターンマッチングの手法を最大限に取り入れたものの一つである．音響・音韻・構文・意味の知識は，有限状態ネットワークに埋め込まれ，マッチングはマルコフモデルと動的計画法を用いて左から右へ順に行われる．

この手法は，音声理解システムで問題となる木探索法・解析手順法・制御法等に関してはできるだけ単純化し，連続（単語）音声の認識手法の拡張でどの程度文音声が認識できるかを占う点で興味がある．つまり，このようなシステムで得られた結果よりも，すぐれた結果が得られるシステムが開発されて初めて音声理解システムの研究が一人前になったといえる．チェスのタスクでは，音響・構文情報を使った場合で約70%，算術文のタスクでは，17% の認識率が得られた．

HARPY システムは，HEARSY-I の開発で得られた経験をふまえて，認識率は比較的高いが計算量の多い DRAGON システムを改良する方向で研究が進められた．両システムで使われていた音響分析（5つのバンドパスフィルターによるピーク振幅と零交差数分析）を LPC 分析におきかえ，分析を精密化した．学習用音声データでシステムを話者に適応させた場合に，ARPA の仕様を満たす結果を得た（1011 の語彙を持つニュース検索のタスクで，5名の話者の184発話文に対して，文の意味理解率95%，文認識率91%，単語認識率 97%．又，算術文の認識率は約 70%，3連続数字音声の認識実験では，話者に適応させた場合で 98%，適応なしだと 93% の単語認識率を得ている）．このシステムが持つ最大の特徴は，ローカル・サーチ（又はビーム・サーチ）と呼ばれる最適な評価を持つパスの他に少数のパスをも並行してサーチする手法で，バックトラックの機能を持たずサーチ時間が早い．この種のサーチのアイデアは，筆者らが既に音声理解システム研究の初期で試みており[18]，以後，バックトラックの機能を持たせた，より一般性のあるサーチ方法を考案している[19]．

プロジェクトの後半3年間の研究主力は，HEARSAY-II システムに注がれた．HEARSAY-I で開発された Blackboard モデルによる独立な各知識源の共同利用法や仮説・検定法を採用している等，文字通り HEARSAY-I の拡張版である．HARPY システムの知識表現が非常に制限された言語モデルに基づいているため単一表現が可能であったのに対し，HEARSAY-II では，種々の独立な知識を用いるために単一表現がとれない等で異なっている．特徴として，① 自己起動や非同期・並列処理を考慮した知識表現法，② 認識処理の途中結果を各階層（パラメータ，音韻，音節，語，構文）軸と時間軸及び候補軸からなる3次元で表現，③ 各階層レベルで，新しい知識の追加が容易，④ α-β 法による探索空間の縮少，⑤ 任意の場所から解析可能，⑥ 並列処理に適したシステム構造があげられる．計画があまりにも膨大であったためか，多大の努力にもかかわらず，途中段階での実験では，認識率・認識時間とも HARPY システムよりも悪く，研究の主力が，現在では画像理解システム等に向けられていることもあって完成の見通しは少ないようである．

なお，CMU では勿論音響レベルの研究も進められたが，目立った成果は認められない．関連研究として，種々のタスクの複雑性の評価法の開発が行われ[20]，システム評価が大切であるとの認識が高まってきたことを示している．

3.3 B B N

BBN では，アポロ計画の月の岩石の化学分析に関する自然言語応答システム（LUNAR システム，語彙数 3500）を語彙数 250 程度の SUS に置き換える研究（SPEECHLIS）から始まった．スタッフは拡張遷移網文法で有名な Woods を中心として，LPC で有名な Makhoul など14名で構成されていた．音響分析と文解析とに，大変しっかりした土台があったが，音声認識の面ではあまり経験がなく，音韻・単語レベルでは，セグメントラティス，単語ラティスと呼ばれる複数個の候補からなる特徴ある出力形式をとったが，ラティ

スへの変換方法に改善の余地があった．SPEECHLISの文解析の特徴は，単語ラティスから seed と呼ばれる最も信頼度の高い単語を中心として解析を進め，セオリーと呼ばれる意味的にまとまった単語例を生成していく意味解析が中心で，構文解析はその結果の検証程度の役割にとどめた点であった．又，インクリメンタル・シミュレーションと称して，一部を人手が介在するシステムで実験を進めながら開発して行く手法をとった事も研究方法に新風を吹き込んだ．

プロジェクトの後半は，ARPA の仕様を満たすべく，1096語の語彙からなる「旅行の計画・予算の問合せ」をタスクとする HWIM (Hear What I Mean) システムの開発にとりかかった．SPEECHLIS との主な違いは，あいまいな単語ラティス上の同定結果に対して，AbS (Analysis by Synthesis, 合成による分析) 手法によるパラメータレベルでの照合を取り入れたこと，不足密度関数と呼ばれる新しい評価関数を導入し優先度の高いセオリーから解析を進めていく手法を厳密化したこと，構文・意味・プラグマティクスの知識をプログラマティク文法に埋め込んだこと，種々の解析手順が比較実験できるように 25 個のフラッグを持つ制御構造を持つことなどである．

解析手順をまとめると次のようになる (図-2 のような代表的な手順と考えられる)．

① 入力発話を線形予測法による分析を行った後，セグメント・ラティスを生成し，これに辞書を走査することによって，解析の核となる単語 (seed) を検出する．スコアの順に event queue に登録する．

② queue から最もスコアの高い event を取り出し，これをセオリーとしてパーサに渡す．

ⓐ セオリーが完全な文なら解析を終了し，意味の解釈結果を返す．

ⓑ セオリーが非文法的なら解析を中断し，それを棄却した後ステップ2を繰返す．

ⓒ その他の場合なら，セオリーに接続可能な単語と意味カテゴリーを予測する．

③ 単語間にわたる音韻論的規則を考慮しながら，予測された単語を同定し，セオリーにその単語を追加したものを新たな event として event queue の該当する位置におく．

④ 新しく生成された event queue に対し，スコアの再評価 (文末の単語の場合)，event の併合 (方向が逆で，かつ，同じ場所で同じ単語を持つ2つの event の場合) 等の処理を行う．

⑤ ステップ2に戻る．

実験は3名の男性が発声した124文章 (1文章は3～13単語) で行われた．約5ケ月にわたって実験とシステムの改良が繰り返され，1カ月に 10% の割合で改善されて最終的に意味理解率が 44% に達した (このことは，非常に細いシステム調整やアイデアが認識率に大きな影響を与えることを示している) が，処理時間が PDP-10, TENEX (PDP-10 のために開発されたページ方式 TSS) で実時間の 1000 倍以上となるなど ARPA の仕様には遠く及ばなかった．

3.4 SRI と SDC

SRI の SUS 研究 (Walker を中心に8名) は，プロジェクトの中でもその研究経験からか人工知能研究としての捕え方が強い．文解析法は，Winograd の自然言語理解システムを変更したものを用いたこともあって，最初の一年間は非常に制限された「積木の世界」を例にとって検討が始まった．ここで，テキスト入力と音声入力との文解析手順の必然的な相違が認識され，以後「積木の世界」よりももう少し複雑な「小型機械器具の組立と修理」をタスク (語彙数は 50 単語程度) として本格的な研究が始まった．解析手順は，左から右へ構文や意味解析においてさえ優先度 (たとえば，WH 型の疑問文の否定形は少ない等) を設けた best-first で行った．単語の同定においてさえ予測された音響特徴が，入力音声中に存在するかどうかを検証するという徹底した top-down 手法であった．このように言語解析レベルでは，ユニークな手法を開発したが十分な実験評価も行われないうちにプロジェクトの前期を終え，結局後期は，音響レベルで実績を上げつつあった SDC との共同研究に移ることになった．

後期は「米・ソ・英の艦隊の船についての情報検索 (SDC で以前から進められていた)」をタスクとして研究が進められた．SRI は言語レベルを分担し，その主な研究内容は，a) システムの効率的制御法，b) システム内の種々の知識源の統合法，c) 新しい意味ネットワーク表現法，d) 文章にまたがる談話内容の管理法であった．一方，SDC は音響レベルを分担し，その主な研究内容は，a) 音響・音韻処理 (処理結果は時間軸とパラメータおよびシンボルのマトリックス表現)，b) 音韻規則辞書，c) 単語整合法(音節整合法も併用) であった．中でも音韻規則の研究は，単語内および単語間共に発音変化の多い英語では重要な研究テーマの一つである．音響分析の特徴の一つは，3つの帯域フィルターによって大まかな分析で有声音区

間を抽出し，その区間に対してのみ LPC 分析を行ったことであろう．

SRI から SDC のシステムへ，ある単語を入力文中で同定して欲しい場所と共に提示すれば，その単語の存在位置と存在尤度が返されるようになっている．SRI の文解析は，初期のシステムと大幅に異なった手法（随時に top-down と bottom-up 戦略の切り換え可能，左右両方向への解析）が採用された．成果として分割型意味ネットワーク表現の開発があげられる．これは，従来の意味ネットワークを分割して，意味内容をネットワーク間表現とネットワーク内表現に分類した．各ネットワーク内では，時間的・意味的に閉じており論理表現における括弧と類似した役割を果たしている．談話のように刻々と現実場面での意味が変わっていく場合も扱え，量記号表現も可能である．しかし，現在の音響レベルはこのような高度な談話理解を必要とするまでには至っていない．SRI, SDC の両システムが完成して間もなく，SDC の計算機システムが解体されたため十分な実験評価はなされなかった．

3.5 その他の研究機関

（a） Lincoln Lab. (MIT)

Lincoln Lab. はプロジェクトの前期で，「音声データの編集」をタスク（語彙数237）とする CASPERS システムを開発した．システム構成は，典型的な階層構造で，文解析は左から右へ bottom-up で行う．音響レベルに重点がおかれており，特徴として単語間にわたる音韻規則，単語辞書に基づく音韻系列の分割・統合の訂正処理を取り入れた単語照合法，言語レベルでは，文脈自由文法の拡張による意味情報の文法への付与，best-first 法の採用があげられる．

プロジェクトの後半は，ARPA との契約は切れたが，独自の研究を続行し，最近では，我々の話す意味内容は一文につき一つであるとの仮定にたち，入力発話文からキーフレーズの抽出研究を行っている．

（b） UNIVAC

音声理解には，韻律情報（ポーズ，ストレス，イントネーション，リズムなど）が有力な手がかりを与えていると議論されていながら，位取り読み数字音声（たとえば nineteen hundred ninety eight）の認識[41]以外まだ本格的に SUS に取り入れられた例はない．これは，韻律と構文・意味とが必ずしも一対一に対応していないこと（時間的にずれる）や韻律情報そのものが不安定であること等によると思われる．

UNIVAC の Lea らは，プロジェクトの契約期間中，音声認識への応用をめざして韻律情報の分析を行ってきた．研究成果として，① ストレスのある音節の認識への利用法，② 発話速度やリズムの自動抽出法の提案，③ 分析・合成手法による構文検出法，④大量の音声データによる韻律現象の調査があげられる．

（c） IBM

IBM は ARPA の SUS プロジェクトとは独自に 1967 年以来連続音声の自動認識に関する研究を進めている．最終目標は音声タイプの実現であり，音声波から音韻系列への変換（FFT（高速フーリエ変換）によるスペクトル分析を使用）と音韻系列から正書法への変換の二つの努力が払われている．一方，SUS プロジェクトに影響されたためか，文章の認識も並行して行われてきた．

音韻系列から単語への変換には，話者ごとに音韻規則や辞書の統計的表現，音響・音韻処理部の統計的なふるまいなどを利用している（このため各話者に対して1時間以上の音声が必要である）．文解析の手順は，初期では左から右に best-first 法で行っていたが，最近では Dragon システムと同じく全パスサーチで行っている．新 Raleigh 言語という 250 語彙からなるタスクを用いた実験では 95% の文認識率が得られた（連続数字音声中の数字の認識率は 98.3% である）．

（d） Bell Laboratory

ベル研究所では，自然言語の連続音声認識の研究はまだ行われていない．最近，区切って発声された自然言語文の自動認識の検討が開始されたところである．本格的な SUS には距離があり，構文情報の有効性の検討や孤立単語音声の認識手法をそのまま文音声に適用しようとしている段階である．

この他，カナダ，イタリア，フランス，西ドイツ，ソビエト等で SUS の研究が行われているが，まだアメリカほどには進んでいない．次章では，我国の会話音声認識（連続音声認識システムというべきかも知れない）研究について述べる．

4. 我国の会話音声認識研究

ARPA の SUS プロジェクトに刺激されて，我国でも文章音声の認識研究が進められるようになり，初期段階の研究発表が 1974 年の電気四学会連合大会で行われた[18]．しかし，連続して発声された日本語文の認識をめざす研究はあまりみうけられず，区切って発声された日本文や人工言語を対象とするものが多く，SUS と呼べる段階に達するものは少ない．

電電公社通研では，列車の座席予約をタスク（語彙数 112）として，オンライン会話音声認識システムの開発が行われた[21),22)]．システムは音響処理部と言語処理部に分かれ，両者はそれぞれ結合された二台のミニコン上で作成され，ソフト的には音韻ラティスを介して結合されている．音響処理部の特徴として，VCV（母音—子音—母音）音節単位によるパターンマッチング法と音韻ラティスをもとに再度音響処理をやり直すフィードバック機構の導入があげられる．単語同定は，単語辞書と音韻ラティスとを種々の音韻変形規則を depth-first 法で適用しながら一致を見い出すことによって行っていたが，棄却される場合が多く最近 fuzzy 的な概念を導入し始めている．言語処理部の最大の特徴は，構文解析手順にも depth-first 法を用いていることであるが，全パスの中から能率よく最良のパスを見い出すことが最近の SUS 研究では多いのと趣を異にしている．しかし，文節ごとに区切って発声するという制限はあるものの，実時間の 5 倍程度の処理時間で 86% の文節認識率が得られている．

京都工繊大では，プログラミング言語 BASIC に若干の制限を加えたものをタスク（語彙数 50）として研究を進めている[23),24)]．意味分析では，ステートメントに付けられた行番号，キーワードの出現順序，変数の定義などの情報を扱っている．解析手順では，depth-first 法と breadth-first 法の両者が文認識率から比較検討されている．

山梨大学では，最初区切って発声されたフォートランプログラムの音声認識の研究を進めていたが[25)]，現在では，連続して発声された日本語文（語彙数 99）の認識システムを開発中である[26)]．音響レベルの処理に重点がおかれているが，言語レベルでの特徴の一つに，一応認識された単語系列をあらためて識別音韻系列とマッチングを行い，単語間にわたる調音結合の影響を再評価している点である．しかし，認識結果がそのまま文の理解につながらない解析手法を用いているという欠点がある．

最後に，日本語音声理解システムの構成，言語レベルの処理手順などの具体例として京都大学で筆者らが開発してきた音声理解システム LITHAN (LIsten—THink—ANswer)[19),27)] について述べる．

本システムは，音響処理部，音韻識別部，単語同定部，単語予測部，解析指示部からなっている．

解析指示部では，述語が末尾に来るという日本語単文の性質と取り扱う世界の仕事の内容を最大限に利用する．日本語単文の場合，述語が決まれば文章構造もほぼ決定するという性質があるから，述語別に文構造を与えることにより，意味情報が容易に扱え（MIT と同様な拡張文脈自由文法を用いる），しかもタスク内で許される文を生成する文法を比較的簡単に構成できる．

まず単語同定部は，そのタスクに関する文章で文解析の上で重要と思われるあらかじめ選定されているキーワードを検出する．解析指示部は，検出されたキーワードの種類とその数に矛盾しない述語群を決定し，それらを入力音韻系列の末尾で同定するように単語同定部を起動する．

単語予測部と単語同定部は，最適と判断された述語に対応する文章構造を種々の知識を使って音韻系列の先頭から順次単語の予測と同定を繰返す．単語の予測の際には，文法ばかりでなくタスクが持つ事実関係の知識も利用する．この場合，予測・同定される単語群は，一般には複数個となることにより多数の単語列が形成される．

文解析を能率よく行うためには，これらの単語列に対して何らかの評価を行って枝刈をし，信頼性のある単語列のみ残して，それに対して以後の解析をする必要がある．そこで，単語の同定結果があいまいであるという性質を考慮した柔軟性のある探索法を開発した．この方法は，信頼性の高い単語列を並行して解析していくもので，best-first 法と breadth-first 法の両者の利点を取り入れたものである．

単語列が信頼できる文章になった時点で解析は終了するが，信頼できる単語列がなくなると次の可能性の高い述語に対応する文構造の解析も始め，両者並行して同様のことを繰返す．

以上の方法により，101 単語の語彙からなる「計算機網への状態問合せと指令」をタスクに選んで実験した．男性話者 10 名が，普通の速度で発声した 200 文章（1 文章は平均 10 単語からなる）に対して 128 文章が正しく認識できた．このときの単語認識率は 93% であった．

このシステムは，文法と単語辞書を変更するだけで，ほかの新しいタスクに適用できる柔軟性のあるシステムであり，実際に「算術文」や「カレンダー」と称するタスクに適用している（孤立数字音声の認識率は，97〜98% である）．LITHAN の主な特徴として，① 他のタスクにも適用可能な柔軟なシステム構造，② パターン理解システムに適したバックトラックの機能を

持つローカルサーチ法，③意味情報やプラグマティクス情報を織り組むことのできる拡張文脈自由文法，④日本語に適した解析手順（キーワードの検出，述語の重視），⑤強力な単語同定法，⑥個人差の教師なし学習機能などがあげられよう．

以上，3章と4章にわたって，会話音声の理解システム研究の現状を概観した．最近の研究方向としては，①perceptual input に適した柔軟性のある木探索法，②発話文の任意の場所からの解析法，③言語レベルからの音響レベルへのフィードバック，④多種類の知識源の制御法などがある．未着手の問題としては，①韻律情報の積極的導入，②間投詞や未登録単語の処理法，③複文・重文の取り扱いなどがあり，今後の研究の進展が期待される．

5. 話者認識研究の現状

5.1 話者認識とは

音声波に含まれる種々の情報の中で，言語情報と共によく研究されているものに話者情報がある．一口に話者認識といっても，与えられた音声が，既知話者集団の一人の発声である場合，その話者を識別する話者同定 (talker identification) と，ある既知話者が発声したものかどうか判定する話者照合 (talker verification) という二つの意味を含んでおり，一般に話者照合の方が容易であるといわれているが，話者同定の問題は，話者照合の拡張としてとらえることができるから両者に本質的な差はない．

話者認識の工学的な応用面としては，主に次のものがあげられよう．

① 話者の確認に用いる場合(話者照合)……口座残高問合せ，チェックレス・バンキングシステム．

② 特定の話者集団の中から話者を決定する場合(話者同定)……個人ファイルの管理サービス．

③ 特定の話者集団の中の人であるかを決定する場合(話者照合と話者同定の併用)……複数人の共有ファイルの管理サービス，ドア・金庫の開閉．

④ 犯罪捜査……特に電話による脅迫犯罪．

さらに，話者認識の研究成果は，機械による音声認識や音声合成に次の点で利用できる．

音声認識……話者による音声パターンの変動の除去．発声者の声質の抽出によって，個人差の正規化あるいは学習に利用できる．

音声合成……聞き手の好みにあった音声の合成．男性・女性・子供らしい声や明るい，つやのある声などの合成に利用できる．

音声における個人性は何に起因するかといえば，音声発声器官の物理的な差（特に声道や声帯の形）や話し癖，方言などである．話者認識の方法は，その手段によって次の三つに大別される．

① 聴き取りによる認識

聴き取りを手段とした聴覚的な判断による方法．

② 読み取りによる認識

声紋 (voice print) に代表されるスペクトグラムなどを視覚的に比較・判断する方法．

③ 機械による認識

人手を介さないで純粋に機械による方法．

一時，ソナグラムの技術開発によって声紋が注目され，指紋や血液型と同様に人を識別する有力な手がかりになると期待されたが，Stevens らによる実験[28]では，聴き取りによる方法が視覚によるよりも話者を識別しやすいとされた．一方，最近の機械による話者認識の研究は，計算機を中心とする情報処理技術の進歩と相俟って音声認識と同様非常な進歩をとげ，ある限定されたタスク（発声内容が既知な場合など）では，機械の方が聴き取りよりも正確に話者を照合できるようになってきた．ここでは特に機械による話者認識について解説する．

話者認識システムを開発するに先だって重要となる仕様は次のようなものである．

① 発話内容に依存するシステムかどうか．任意の発声内容から話者認識するには，音韻要因と交互作用のない個人性パラメータを抽出するか音韻認識をする必要がある．

② 登録用音声の発声時とシステム使用時の時期差は，どの程度まで許すか．個人性パラメータは，発声時の時期差によって大きく変動することが知られている．

③ 発話の時間長をどの程度要求するか．一般に発話内容が長くなればなる程認識しやすくなる．

④ 周波数制限を設けるかどうか．話者認識の応用面では，電話回線を使用する場合が多くなると思われるが，その周波数制限 (300～3400 Hz) に対処できうるかどうか．

⑤ 人手の介入を許すかどうか．犯罪捜査などのように，認識に要する時間よりも結果の信頼性に重きを置く場合は，man-machine interaction の導入が有望である．

結局，与えられた音声から上述の仕様を満たす特徴

パラメータを抽出することが研究課題となっている．次節では，主に話者を特徴づける個人性パラメータとその経時変化についての最近の研究動向を述べる．紙面の都合により実際の認識アルゴリズムやシステムの現状については述べることができないので文献を参照されたい[29)~32)]．

5.2 個人性を表わす特徴パラメータ

（a） 特徴パラメータ

ある特徴パラメータが話者認識に有効かどうかの評価は，正確には話者認識実験に待つよりほかないが，通常，話者間での分散に対する話者内での分散の比が大きい (F-ratio) ことを評価基準としている．さらに考慮すべき点として，抽出が容易なこと，経時的に安定なこと，音声中によく現われるものであること，話し方や話者の健康状態等によってあまり変化しないこと，他の話者によってまねにくいものであること，周囲雑音に強いこと等があげられる．

個人性パラメータとしてよく用いられているものには，音圧（ゲイン），基本周波数，短時間スペクトラム，フォルマント周波数や帯域幅，ケプストラム，線形予測係数，PARCOR 係数，各音韻の相対的な持続時間などがある．これらの多くは，音声認識に関しても使われているものである．

Wolf は，基本周波数や母音・鼻音の各スペクトル，フォルマント周波数，音源スペクトル，発話時間長などの特徴パラメータから，個人性をよく表わしている特徴を求めた[33)]．それによると，コンテキスト別の基本周波数が最も話者を特徴づけるパラメータであること，母音と鼻音とでは特に差はなく，第2フォルマントの方が第1フォルマントよりもすぐれていると報告されている．

Sambur も同様な研究を行った[34)]．特に分析対象話者の3年半にわたる音声を使用した事など Wolf よりも厳密な分析を行い，/n/ の第2フォルマント，/m/ の第3・第4フォルマント，母音の第2～4フォルマントが，基本周波数では，その変化の傾斜よりも平均値が個人性を特徴づける有効なパラメータだとしている．

又，Rosenberg と Sumbur は，フォルマント周波数と線形予測係数の比較を行い，後者の方がすぐれていると報告している[35)]．

Atal は，全極モデルより導出される線形予測係数，全極フィルターのインパルス応答，自己相関関数，声道断面積関数，インパルス応答のケプストラムの相互比較を行った[36)]．発声された文を 40 分割し，上述の各パラメータを抽出したものに2次判別関数を適用して各区間ごとに話者同定を行った．それによるとケプストラムが一番すぐれており，次が線形予測係数で，声道断面積関数が最も悪かった．

上述の特徴パラメータは，基本周波数を除いて，すべて音韻に依存しており，又，音韻と話者の交互作用もある．発話内容が決まっている時は，音韻要因を取り除き話者要因だけを抽出できようが，発声内容が不定の場合には，あらかじめ音韻認識を行うか，あるいはそれと類似な処理をする必要があり，音韻に依存する特徴パラメータを用いる方法は非常に難しくなる．そのため発声内容に依存しない話者認識では，各音韻に共通に含まれる特徴を用いる方法がとられる．中でも，長時間にわたる平均的なパラメータ（たとえば，長時間平均スペクトル）を用いることが多い．しかし，この方法は，発話サンプルが十分長い時のみ有効で，発話サンプルが短い場合は，どうしても音韻認識を介する方法をとらざるを得ない．Markel らは，スペクトルと基本周波数，ゲインのそれぞれの長時間平均による話者認識実験によって，相互比較を行い，スペクトルの長時間平均が最もすぐれていると報告している[37)]．

田畑らは，近似的に音声を声道特性と声帯特性に分離し，どちらにより多くの個人性情報が含まれているかを調べた[38)]．それによると，声帯特性よりも声道特性により多くの個人性情報が含まれているようである．同様な主旨の研究が，伊藤らによって聴覚実験を通して進められている[39)]．

いづれにしても興味ある問題であるが，ただ声道形の抽出が完全にはできないことや聴覚に有効なパラメータが機械にとって有効かどうか（その逆も含めて）の疑問もあり，今後の研究に期待されるところが大きい．

（b） 特徴パラメータの経時変化

最近の話者認識研究の成果の一つに，登録用音声の発声時と認識用音声の発声時の時期差が大きくなると，特徴パラメータも大きく変わることが定量的に明らかになったことがあげられる[40)]．これらの研究から，時期差が1カ月以上になると急激に話者認識が難しくなること，しかし，長期間にわたって発声された音声によって作成された標準サンプルを用いれば，数カ月以上の時期差でもかなりの話者認識率が得られることがわかった．又，近似的に音声を声道特性と声帯特性に分離して，個人性パラメータの経時変化を分析

したところ，声道特性よりも声帯特性の方が経時変化が大きいことも明らかになっている．

この他に，同一話者の個人性パラメータの変動要因としては，文脈の違いによるもの，ピッチの違いによるもの，心理状態の違いによるもの，健康状態（風邪など）の違いによるものなどがあり，これらをどのようにして取り除くかが今後の研究課題となろう．

6. おわりに

最近10年間の音声認識や話者認識研究の進歩は目覚ましいものがある．これは，電子計算機を中心とする情報処理技術の進歩に負うところが大きい．以前の研究には，少ない音声データに基づく報告が多かったことや追試が容易でなかったこともあってその信頼性に疑問があり，結局研究成果の蓄積とならなかったきらいがある．最近では，大量の音声データで厳密な実験を行うことができるようになったため，開発された認識アルゴリズムの評価の信頼性が高くなり，研究成果が着実に蓄積されてきている．

会話音声の認識・理解研究は，ようやく研究成果が現われてきた時点であり，ARPA の SUS プロジェクトの解散によって下火となってきたのは残念である．工学的応用面でも人工知能面でも興味ある研究テーマであり，今後とも地道な研究努力が望まれる．ただし目標が大きすぎたことや音響レベルの軽視が ARPA の SUS プロジェクトを解散に追い込んだことを反省する必要がある．ARPA の仕様を満たさずとも，100～200 語彙程度からなる実用的なタスクが多く存在すると思われるからタスクをうまく設定すれば，自然言語による計算機への音声入力が可能となろう．又，音声認識の応用場面では，機械との対話形式を取ることができると思われ，少々認識や理解の誤りがあっても，対話的に誤訂正できる機能を付加しておけばあまり問題とはならないだろう．しかし，話者認識の応用場面では，このような設定はむしろ例外と考えられ，工学的応用への実現には，より一層の研究努力が望まれよう．

筆者の不勉強で，音声認識と話者認識の研究現状を概観したにとどまり，読者に今後の研究方向を伝えただけの十分な解説ができなかったことをお詫びしたい．

最後に，本稿をまとめるにあたって貴重な御教示を戴いた坂井利之教授をはじめ音声研究グループの皆さんに感謝致します．

参考文献

1) P. Denes: The Design and Operation of the Mechanical Speech Recognition at University College London, Jour. Brit. I. R. E. Vol. 19, No. 4, pp. 219～234 (1959),
2) S. Doshita: Studies of the Analysis and Recognition of Japanese Speech Sounds, 京都大学学位論文 (1965).
3) R. Alter: Utilization of Contextual Constraint on Automatic Speech Recognition, IEEE Trans. Vol. AU-16, No. 1, pp. 6～11 (1968).
4) J. McCarthy et al.: A Computer with Hands, Eyes, and Ears, Conference Proceedings of AF IPS, Vol. 33, pp. 329～338 (1968).
5) G. Fant: Automatic Recognition and Speech Research, STL-QPSR 1, pp. 16～31 (1970).
6) D. R. Reddy: Speech Recognition: Properties for the Seventies, Proceedings of IFIP, pp. I5～13 (1971).
7) A. Newell et al.: Speech Understanding Systems: Final Report of a Study Ggroup, North-Holland (1973).
8) Proceedings of International Joint Conference on AI (1973, 1975, 1977).
9) L. D. Erman, Ed.: Contributed Papers of IEEE Symposium on Speech Recognition, Carnegie Mellon University (1974).
10) A Special Issue on Speech Recognition, IEEE Trans. Vol. ASSP-23, No. 1 (1975).
11) D. R. Reddy, Ed.: Speech Recognition, Academic Press (1975).
12) Conference Record of International Conference on ASSP (1976, 1977, 1978).
13) D. R. Reddy: Speech Recognition by Machine: A Review, Proceedings of the IEEE, Vol. 64, No. 4, pp. 501～531 (1976).
14) R. De Mori: Advances in Automatic Speech Recognition, Proceedings of the 4-th International Joint Conference on Pattern Recognition (1978, to appear).
15) D. R. Reddy: Pragmatic Aspects of Machine Vision, Dept. Inform. Science, Carnegie-Mellon University (1977).
16) D. H. Klatt and K. N. Stevens: On the Automatic Recognition of Continuous Sspeech: Implication from Sspectrogram-reading Experiment, IEEE Trans. Vol. AU-21, No. 3, pp. 210～217 (1973).
17) M. F. Medress et al.: Speech Understanding Systems, Report of a Steering Committee, SIG ART, Newsletter, No. 62, pp. 4～8 (1977).
18) 音声認識における言語情報の利用，電気四学会連合大会シンポジウム (1974).

19) T. Sakai and S. Nakagawa: A Speech Understanding System of Simple Japanese Sentences in a Task Domain, 電子通信学会論文誌, Vol. E 60, No. 1 pp. 13～20 (1977).
20) R. G. Goodman: Analysis of Languages for Man-machine Voice Communication Ph. D. Dissertation, Stanford University (1976).
21) 鹿野, 好田: 会話音声の機械認識における言語処理, 電子通信学会論文誌, Vol. 61-D, No. 4, pp. 253～260 (1978).
22) 中津, 好田: 会話音声の機械認識における音響処理, 電子通信学会論文誌, Vol. 61-D, No. 4, pp. 261～268 (1978).
23) 新美, 浅見: 音声認識システムにおける言語情報の利用とその効果, 電子通信学会論文誌, Vol. 58-D, No. 12, pp. 741～747 (1975),
24) 新美他:「Spoken BASIC 1」の認識システム, 情報処理, Vol. 18, No. 5, pp. 453～459 (1977).
25) 関口他: フォトラン・プログラムの音声認識システム, 情報処理, Vol. 18, No. 5 pp. 445～452 (1977).
26) 関口, 重永: 日本語文章の音声認識システム, 日本音響学会誌, Vol. 34, No. 3, pp. 204～213 (1978).
27) 坂井, 中川: 不特定話者・連続音声向き単語音声の識別, 情報処理, Vol. 17, No. 7, pp. 650～658 (1976).
28) K. N. Stevens et al.: Speaker Authentication and Identification: a Comparison of Spectrographic and Auditory Presentations of Speech Material, JASA, Vol. 44, No. 6, pp. 1596～1607 (1968).
29) B. S. Atal: Automatic Recognition of Speakers from Their Voices, Proceedings of the IEEE, Vol. 64, No. 4, pp. 460～475 (1976).
30) A. E. Rosenberg: Automatic Speaker Verification: a Review, Proceedings of the IEEE, Vol. 64, No. 4, pp. 475～487 (1976).
31) 中田: 音声, 日本音響学会編, コロナ社 (1977).
32) 古井: 話者認識研究の現状と問題点, 視聴覚情報研究会資料 (1977年6月).
33) J. D. Wolf: Efficient Acoustic Parameters for Speaker Recognition, JASA, Vol. 51, No. 6, pp. 2044～2056 (1972).
34) M. R. Sambur: Selection of Acoustic Features for Speaker Identification, IEEE Trans. Vol. ASSP-23, No. 2, pp. 176～182 (1975).
35) A. E. Rosenberg and M. R. Sambur: New Techniques for Automatic Speaker Verification, IEEE Trans. Vol. ASSP-23, No. 2, pp. 169～176 (1975).
36) B. S. Atal: Effectiveness of Linear Predictive Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, JASA, Vol. 55, No. 6, pp. 1304～1312 (1974).
37) J. D. Markel et al.: Long-term Feature Averaging for Speaker Recognition, IEEE Trans. Vol. ASSP-25, No. 4, pp. 330～337 (1977).
38) 田畑, 亀井, 大野: 母音における話者要因の聴取評価——声道特性と声帯特性の分離, 日本音響学会講論集 (1977年4月).
39) 伊藤, 斎藤: 個人性の知覚に寄与する音響パラメータの分析, 日本音響学会講論集 (1977年10月).
40) 古井: 音声の個人性パラメータの時期変動と話者認識, 電子通信学会論文誌, Vol. 57-A, No. 12, pp. 880～887 (1974).
41) Y. D. Willams: The Use of Prosodics in the Automatic Recognition of Spoken English Numbers, Ph. D. Dissertation, MIT (1972).
42) D. H. Klatt: Review of the ARPA Speech Understanding Project, JASA, Vol. 62, No. 6, pp. 1345～1366 (1977).
43) W. A. Lea and J. E. Shoup: Gaps in the Technology of Speech Understading, in 12) (1978).

(昭和53年3月8日受付)
(昭和53年4月27日再受付)