

第五章 音声認識技術の成長

一、音声認識研究の始まり

言葉を話す行動をコンピュータに行わせる「情報発信系」は比較的容易であったが、聞こえてくる言葉を受け取る「受信系」は、人間が介在しても、コンピュータに行わせることは、いまだに難しい。

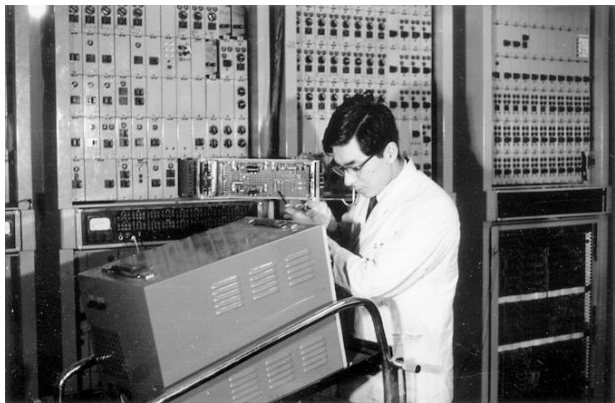


京大における初期の音声認識研究

音声認識技術の研究は、一九五二年、アメリカのAT&Tベル研究所で、ゼロ交差回数法というゼロ軸を通過する音声波形の回数を数えて周波数を検出する手法を用いて、数字の認識を行う報告が行われたときから始まった。この報告を基に、京都大学教授の坂井利之は、ただ単によい品質で効率的に音声を送送する技術としてではなく、音声を情報処理の対象として位置づけようとした。一九五九年、助手の堂下修司(写真Ⅱ白衣の男性)とともに日本電気と共同でベル研の成果を拡張した単音節認識装置「音声タイプライター」の開発に乗り出した。一九六〇年七月発売の「科学朝日」にその試作品が紹介された。「ワ」「タ」「シ」と一音ずつ発音すると、一文字ずつ打たれるというものであったが、これが音声認識技術の第一世代の始まりであった。一九六二年、西独ミュンヘンで開かれた情報処理国際連盟(IFFIP)で坂井は胸を張って、音声タイプライターの研究成果を報告した。

音声認識の萌芽期とともに歩み出した日本電気の迫江博昭は、地道に研究を進めていたが、一九六八年迫江は、DP(ダイナミックプログラミング)マッチング法を開発した。これは、一九七一年、「連続音声認識装置」(願S461074828、特0986698号)の名称で出願され特許が降りた。

明細書の特許請求範囲には、自然な人間の話の流れにつれて、言葉の標準パターンの中で最も確かな候補を選び出して決定していく回路の構成と機能を説明している。この技術は、主に特定話者の音声認識に用いられ、話された音声の特徴と登録してある音声の標準モデルとを比較して照合する方法で、発声速度が揺れても一定のパターンを検出できるように工夫されたことから、DTW(ダイナミック・タイム・ワーピング)法とも呼ばれ、多くの研究者によって二段DP法など、機能を改善するための研究が続けられた。一九七八年には日本電気は世界初の連続音声認識装置「DP100」を発売、荷物の配送先仕分けなどに普及した。



このD P マッチング技術は、同時期にロシアのビンツュック Vintsyuk も報告していた。動的計画法と呼ばれ、日本が主導権をとった、この時期が第二世代に当たる。

第三世代は一九八〇年からの連続音声認識の始まりであった。アメリカではカーネギーメロン大学（CMU）、IBM に続いて、ベル研で隠れマルコフ法の研究が進められた。IBM が発明した隠れマルコフモデル（HMM）という理論が今日の音声認識技術の主流となる。不特定話者の多量の音声データから得た特徴パラメーターの時系列パターンから「最尤推定の原理」（EM アルゴリズム）に基づいてHMM による音響モデルを作成し、ビタビアルゴリズムにより入力音声の特徴パラメーターの時系列パターンの尤度（確かさ）を計算するものである。言い換えれば聴取可能な音声記号の列から、変化していく状態における言葉の出

現確率分布に応じて、確からしい文字列を導き出そうという手法である。それまでの方法では、いろんな人の音声を想定すると音声の特徴を示すバリエーションがふえて、普通のD P 法やパターンマッチング法では処理できないので、数学に基づく統計的な枠組みから自然な話し言葉を認識させる技術として研究され、ベル研究所が普及の役割を担った。不特定話者の音声処理する汎用技術となるものであった。この世代から、研究開発の主役はアメリカに移った。

といっても、この時期に我が国でも音声認識研究が非常に進んだ。国際電信電話の山本誠一が新たに発足した国際電気通信基礎研究所において音声翻訳研究に乗り出していた。その最新成果は、スマートフォンなどに組み込まれている。また、奈良先端大学の鹿野清宏はいち早くHMMを導入し、海外研究者と連携を進めた。さらには、文節で区切って入力していたモデルから、今日の連続音声認識への土台が豊橋技術科学大学の中川聖一によって築かれていった。

一九八〇年代からは、数学のニューラルネットワーク（NN）を用いたパターン認識が盛んに研究された。しかし、HMMを超える認識性能が確認されないため一九九〇年代前半には停滞している。

第三世代の枠組みを越えていないため現在はまだ三・五世代と呼ばれているが、一九九五年には名実ともに連続音声認識が可能になってきた。その担い手は、アメリカ政府が一九九〇年代に巨額の資金を投じて行われたDARPA（国防高等研究計画局）プログラムだった。

アメリカでは国を挙げて基礎研究から着手し、電話音声、自然会話など徐々に高度な研究へ進んだ。参画したのは、ベル研究所、IBM など、大学では、CMU、マサチューセッツ工科大学（MIT）、オレゴン科学技術大学院大学（OGI）など、研究機関ではスタン

フォード研究所などであった。その中から、スピーチワークス社、ニュアンスがベンチャーとして誕生し、電話系の音声認識ビジネスを育てていくことになる。

一九八〇年代後半から一九九〇年代前半にかけて、DARPAはディクテーション（口述筆記）プロジェクトを実施した。研究の成果として、文章の言語モデルとしてNグラムモデルが提案された。これが大語彙連続音声認識につながっていった。Nグラムというのは、統計的自然言語処理分野で広く使われている言語モデルで、連続するN語間の関連データを確率的に計算処理して連続する音声を識別する手法である。



河原達也

一方、日本では京都大学河原達也（写真）、奈良先端大鹿野らが、音声認識システムの開発、研究者のためにオープンソースの高性能な、汎用大語彙連続音声認識エンジン「ジュリアス」（JULIUS）を一九九七年から開発、公開してきている。これは、LINUX、マック、ウインドウズその他iPhonなど、さまざまなプラットフォームで動作する汎用性が高い認識エンジン。またオープンなア

ーキテクチャーを採用しており、認識に用いられる音響モデル、単語辞書、言語モデルを明確に分離しているため、日本語だけでなく外国語のアプリケーションにも移植されている。数万語に及ぶ大語彙量の連続音声認識をパソコン上でほぼリアルタイムに実行できる性能があり、ソースコードも無料で入手できるように配慮された。ジュリアスに発音辞書や言語モデル、音響モデルなどのモジュールを組み換えると、さまざまな用途に応用できるものであった。この技術公開が日本の音声認識技術の進歩に大きな貢献を果たしていくことになる。

二、アメリカではステノマスクからボイスライターへ進化

文字書き技術は、文字文明が誕生して以来、人間が自分の手で筆記することであった。しかし、機械文明、コンピュータ文明は少しずつ手書き文化を書き換えている。速記という文化も文字と同じような経過をたどってきたが、早書き技術という性質上、文字よりも技術革新の波が早く押し寄せてきている。

アメリカでは機械速記がその性能のよさゆえに、手書き速記界から認められない時期を味わったが、今日では機械速記が速記の代表選手になっている。



一方、第二次世界大戦が始まってから一九四〇年代、シカゴの手書き速記者、ホーレイス・ウェブ Horace Webb が「ステノマスク」を考案した。法廷で速記符号を書くかわりに、消音器をつけたマイクروفोनを使って、聞こえた発言を復唱しながらテープレコーダーで録音する新しい発言記録の作成法であった。

録音された音声は、タイピストがトランスクリプションとして

完成する仕組みで、この反訳分業化はテープレコーダーが登場して以来、機械速記者が用いていたから、音声固定工程を録音技術の普及で第三の方法である「復演」という手段によって解決したわけである。もっとも、この分業体制はアメリカだけでなく、かつて日本でも教育や人海戦術として復演という名称で行われていた。

アメリカでは全米速記者協会（NSRA）時代から、毎年速記チャンピオンを競うコンテストが開かれているが、ステノマスクの速記者たちの全国団体である全米逐語速記者協会（NVRRA）でも一九七二年からナショナル・スピード・チャンピオンシップが開かれてきた歴史がある。分速三百五十語の超高速に対応しており、一九九四年から三年連続で優勝したホルト女史 Dian C. Holt は、一九九六年に正確度九九・二〇％を挙げた。コンピュータ文明の所産、CATは一九八〇年代に機械速記に技術革新をもたらした。音声認識に関しては、アップル社が一九八九年、ボイス・ナビゲーターを商品化し、三年後の一九九二年にIBMがスピーチ・サーバーを発売していたが、ステノマスクの分野には一九九七年にドラゴン社のナチュラリー・スピーキング、同年末IBMのピア・ボイスが認識エンジンとして登場した。

今日、ステノマスクはボイスライターと呼ばれているが、装備の構成は遮音型マイクのステノマスクのほか、音声デジタル化デバイス、パソコン、音声認識エンジン、CATソフトとなっている。使用される音声認識エンジンは、DARPAの成果の一つであるドラゴン・ナチュラリー・スピーキングである。

ボイスライター（写真）の長所は、人間が本能的に備えている話すという機能をそのまま利用する点にあり、速記符号を記憶して熟練するという過程を必要としない。聴解能力と表記能力、勘のよさと反応の速さが素質として備わっているオペレーターに適した手法である。初心者には二年間、法廷などでの議事録作成作業に従事している秘書などでは一年以内でリアルタイム作業が可能となるという。ただ、ボイスライターは古い世代の特定話者による音声認識技術であり、現在の不特定話者認識ではないことが課題である。音声認識は句読点を打たないことや、訂正作業が能率的でないことなど、機械速記との質の競争において解決すべき点がある。



G. P. トリブルツィオ

三、欧州における会議録作成の状況と音声認識の普及

機械速記と同様、英語よりも、イタリア語のほうが音声認識技術の登場は早かった。欧州における国家的プロジェクトは一九八四年からESPRITが開始されているが、二〇〇〇年代からオランダ、フランス、ドイツ、スペインでも普及が始まっている。世界的に見れば、米日欧入り乱れて、エアバス、ボッシュ、ドコモ、エレクトロラックス、ホンダ、IBM、インテル、メルセデス、パナソニック、フィリップ、ソニー、スキャンソフト、ノキア、ジーメンスといったグローバル企

業が数十億ドルを投じて音声認識技術の産業応用を進めてきている。

ヨーロッパ・コミュニティ（ＥＣ）は、五十三カ国から構成され、アメリカと同じぐらいの人口があり、実に二十三言語が話されている。そのため域内コミュニケーションが大きな課題だが、既製音声認識ソフトが利用できないため、音声認識の利用は進んでいない。これまでＥＣは幾つかの調査研究プロジェクトを支援してきたが、日本のジュリウス、スフィンクスなど、音声認識のオープン・ソース（無料）ソフトの応用にも期待されている。

○欧州議会（ヨーロッパ・パリアメント） 音声認識は、オリジナル言語で使用された文書を翻訳する通訳によって、本会議の会議録を作成するために使用されている。議事録担当者はニュアンス社のドラゴン・ナチュラリー・スピーキング・プロフェッションナル版（ＤＮＳＰ）を使用している。ドラゴンは、オランダ語、イタリア語、フランス語、ドイツ語、英語版がある。

○チェコ共和国 ニュートン・インフォメーション・テクノロジー社がリベック・工科大学と共同でニュース配信ネットワーク用にチェコ語の音声認識を研究している。同社は、アメリカのブルームバーグ社やトムソン・ファイナンシャル・ネットワークと同じように幅広く産業別、部門関連別ニュースを電子配信している。

○英国 英国はテレビ放送の字幕分野で高い目標を設定している。ＢＢＣは二〇一〇年を目標に新しい番組に一〇〇％の字幕付与を目標としてきた。現在、五十人のスタッフメンバーの半分近くが字幕つけに音声認識を使用している。年間三千時間から四千時間の字幕がオーストラリアの機械速記者を用いて、ワーキングによりサービスされている。

イギリス議会の下院では、速記者が録音データから書き起こす作業形態を続けており、校閲者が議場でデジタルメモをとって、速記者へ渡す。音声認識の利用はまだ行われていない。

○ドイツ ハノーバー議会では二〇〇二年音声認識を導入して、会議録作成の速記に置き換えた。法廷記録では二〇〇五年初頭までに、約三千台のコンピュータにニュアンス社のドラゴンソフトが装備された。音声認識は主に裁判官や秘書のワークフロー管理と呼ばれる事務手続きに利用されている。ステノマスクは現在行われていない。

○イタリア 国会の上院では一九九六年からＩＢＭのドラゴンを用いて復演する方式を導入している。本会議の会議のメモを普通文字で書き、復演の際に利用する。要約記録を作成する作業に議事録担当者が従事している。上院では十年間をかけ、音声認識を用いて最終文書を作成するために準備してきた。下院では二〇〇〇年から逐語会議録はパソコンで書き起こしにより作成されている。これら議事録担当者たちの執務は、議場で二十分間記録メモをとり、その後、ワークフロー管理によって統合された音声認識ソフトを用いて音声記録から書き起こす。

地方議会のシチリアナ・パレルモでは、委員会の要約報告を作成するために音声認識が使われている。下院と同じように、会議中の議論のために音声認識を導入する計画がある。他の二十一地域からの報告では、州や市議会本会議はデジタル録音し、音声認識を用いて

文書化しており、文書化を要める作業量はふえている。

裁判では現在のところステノマスク技術は使用されていない。二〇〇六年十一月時点で約六〇％の法廷調書の反訳がデジタル録音から行われ、四〇％が機械速記で作成されている。

大学や学校では、多くの教師がこの技術を障害学生の支援に使いたいと願っており、大学では主にステノタイプが使われているが、近い将来音声認識が導入されるだろう。

ライ（RAI）というイタリア全国ネットのテレビ字幕つけのうち、オフラインの字幕付与に音声認識が用いられている。リアルタイム字幕の教育訓練がボローニャ大学などで試みられている。

○ベルギーとデンマーク アントワープ大学とコペンハーゲンビジネススクールは、読みやすさの改善、字幕合成の研究に関連した教育的プロジェクトを研究している。

○オランダ オランダ国会はデジタル録音を含む音声認識会議録作成法「VLOSシステム」を採用している。閣僚、議員等発言予定者と議題、動議などをメニューに登録しておき、発言記録を担当する臨場者がメニューをクリックして必要な情報を記録する。それぞれの発言の音声ファイルにその情報をつけて、逐語記録作成者へ渡して会議録を作成する。京大、河原によれば「実現はまだだが、会議録と映像をリンクしたアーカイブの構築が可能で、各国の議会関係者から関心を寄せられている。

○フランス ビジュアル・ボックス社は、ステノタイプか音声認識を用いるオーケストラシステムを開発、リアルタイム出力がインターネットのブラウザを使って直ちに訂正、更新される。

フランス議会下院では、二十一世紀に入って速記者の養成を停止した。現状では記録作成者が十五分交代で議場で記録メモをとり、その後音声ファイルを基にワープロを用いて逐語記録を作成する形態をとっている。音声認識はまだ導入されていない。

○スイス 四公用語があり、国内テレビにはそのうちイタリア語、フランス語、ドイツ語の三言語について音声認識を用いて字幕がつけられてきた。スイスはほかの放送にも拡張しようとしており、この技術を使って大学生の教育にも見通しを持っている。

○トルコ 二〇〇〇年六月、コック・ホールディング・インフォメーション・テクノロジ―社グループのGVZ社は、トルコ語音声認識と読み上げ（TTS）技術を開発した。

○ロシア ロシア語における音声認識研究は、大学、国立研究所、民間会社の産官学連携で進んでいる。GZOS速記訓練国立センターは、英語、ロシア語の音声認識を訓練している。

○ポーランド ポーランド議会では速記者は議場で記録メモをとり、議事録作成担当者が録音からパソコンで書き起こし、校閲者が修正する工程をとっている。音声認識は採用されていない。

（注）迫江博明の連続音声認識装置」の明細書には「複数個の単語からなる語彙の中から数個の単語を連続して発声した入力音声の特徴ベクトルの系列として表現して保持する機能を有する入力パターン保持部と、上記語彙に含まれる各単語に対応する標準パターンを保持する機能を有する標準パターン記憶部とを備え、標準パターンの特徴ベクトルと入力パターンの特徴ベクトル間の類似性を評価する機能を有するベクトル類似度計算部と、複数個の数値を記憶しシフトする機能を有する第1および第2の演算レジスタおよび第1最大値選択回路と加算回路により本文中（2）式に示すような動的計画法の計算を実行する手段を構成し、かつ（4）式の計算を実行する正規化部と第2最大値選択部によって音声パターンの類似度と最大値位置を決定し、各標準パターンに対して計算される類似度が最大である標準パターンを決定し認識結果を出力する決定部を有し、その標準パターンに関する動的計画演算の最終結果である第1演算レジスタの内容を一時保持するための第8演算レジスタと、第8演算レジスタの内容を、類似度が最大である標準パターンに対して決定される最大値位置の情報をもとにしてシフトする手段を有することを特徴とし、入力パターンの次の単語を認識するにあたって、第8演算レジスタの内容を第2演算レジスタに移しそれを初期値として動的計画演算を行い類似度を計算することにより入力音声を順次認識することを特徴とする連続音声認識装置」と書かれている。