3．特集：音声認識・合成と自然言語処理の研究開発動向 －人に優しいヒューマンインタフェース実現への課題－

情報通信ユニット 亘理 誠夫

3.1 はじめに

音声認識合成技術や自然言語処理技術は人が自然な形で機器を使用するための入出力技術として昔から研究されてきた。人が機器を使うためのヒューマンインタフェース技術としては、初期のコンピュータでは、テキストによるコマンド入力とコンピュータからのテキストによるメッセージ出力であった。その後、アイコン表示とマウスによる選択というグラフィカルなインタフェースとなった。また、コンピュータグラフィックスの進歩、画像・音声・オーディオなどマルチメディア処理の進歩により多彩なインタフェースが出現している。さらに、使い勝手のよさを向上させるため、画面のデザイン、多種多様な入力デバイスの研究も進められている。

しかし、依然として情報機器のヒューマンインタフェースはある程度の習熟を要求し、人が人とコミュニケーションするように自然な形で簡単に使用できるまでに至っていない。例えば、音声や自然言語（人が通常コミュニケーションで使っている言葉）で情報システムと対話して、情報を収集したり、場合によっては自動翻訳により外国語の情報を入手したりすることが、究極の姿であろう。

このような人にとって自然な音声や自然言語を用いるインタフェース技術は、古くから研究され発展し、限定された範囲では利用されるようになってきた。ワープロへの音声入力(ディクテーション)や Web 上の外国語の情報を読むための粗いが簡便な翻訳などが実現している。しかし、まだ通常の会話音声の認識率は低く、また翻訳文の品質が高くないという課題が残されている。

一方、インターネットの広がりと共に、PCや携帯電話、携帯情報端末など情報機器が広く普及しており、コンピュータを使い慣れた人だけでなく、初心者、高齢者を初め「だれでも」「どこでも」「簡単に」使える情報機器のインタフェースが強く望まれるようになってきている。総合科学技術会議においても、次世代のブレークスルーをもたらす基礎的、萌芽的な領域の研究開発として、「機械が人に合わせて高度なコミュニケーションができる意味理解技術等のヒューマンインタフェース技術」を10年後に実現することを目指している。

本報告では、音声と自然言語を用いるインタフェースの研究開発動向について、研究の発展と現状、日米における研究プロジェクト推進方法の相違を述べ、次世代ヒューマンインタフェースの研究を促進するための課題提言を試みる。

3.2 ヒューマンインタフェース技術の発展と現状

3.2.1 音声認識

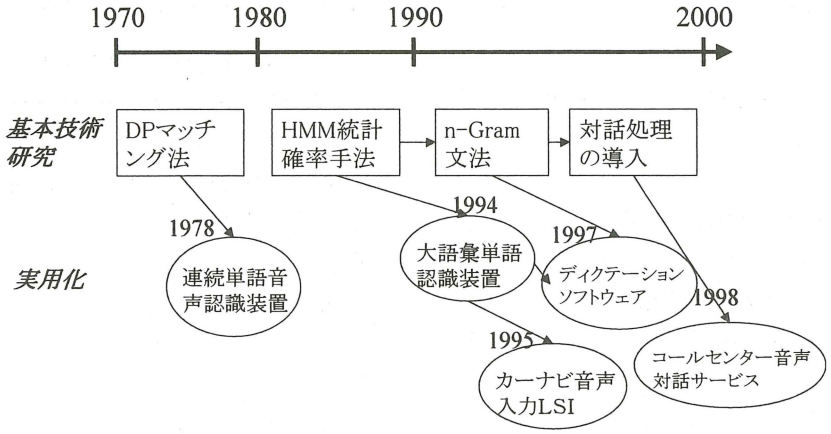
(1)発展の歴史

人が発声した言葉を機械に認識させるという音声認識研究の歴史は、1952 年ベル研究所の Davis らによるゼロ交差数[1]を用いた数字音声認識の試みから始まった。1959 年に京大においてそれを拡張した単音節認識装置「音声タイプライタ」が研究された。実用化につながるブレークスルーとしては、1970 年代に発声時間長の伸縮を動的計画法を用いて正規化するDPマッチング法[2]が日本とロシアで同時に提案され、さらに日本で連続数字を認識できる2段DPマッチング法が提案された。1978 年にこの方式を用いたミニコンベースの連続単語認識装置が日本で製品化され、荷物の仕分けなど手がふさがれている状況でのデータ入力に使われた。

米国では、1970 年代に統計確率的手法であるHMM(Hidden Markov Model)の研究が行われ、1980年代には単語音声認識の標準的手法となった。1980年代後半から 1990 年代前半にかけて、DARPA（国防省高等研究計画局）がディクテーションプロジェクトを実施した。これにより文章の言語処理に n-Gram 法(n語間の統計確率を用いる方式)が提案され、大語彙連続音声認識が実現された。この成果をベースに、PCの性能の向上を背景に、ディクテーション用の大語彙連続音声認識PCソフトウェアが 1997 年米国で販売されるようになった。日本でも同時期に日本語用のソフトウェアが発売された。

また、1990 年代前半からは、DARPA プロジェクトにより、Q&A対話システムの研究が行われ、Q&Aをコントロールするための対話処理技術の研究が開始さ

図表 1　音声認識研究と実用化の流れ



れた。その成果をベースに米国では、1998 年に、電話による各種予約・問い合わせサービス(コールセンターサービス)の自動音声対話処理が実用化されている。

一方、日本で実用化研究が進んだのは、カーナビの音声入力である。自動車内で地名やコマンドの音声入力を可能とするため、認識性能を落とさず各種処理を軽減したアルゴリズム開発が行われ、1995 年に実用化している。

(2)現状と課題

現在の音声認識の技術レベルは、明瞭な発声で読み上げた文章は、概ね正しく書きおこしができるまでになっている。また、カーナビなどで利用されているようにある程度の騒音下でも認識はできている。

しかしながら、考えながら発声した「話し言葉」や、知人同士の「対話音声」については、まだ認識性能は低い。また、対話において、発声者の意図を読み取ったり、状況を判断することはできていない。

3.2.2　音声合成

(1)発展の歴史

音声認識と対をなす音声合成、すなわち、機械からメッセージを音声によって出力する研究も 1950 年代に始まっている。MIT(マサチューセッツ工科大学)の Stevens、KTH(スウェーデン王立工科大学)の Fant らにより声道音響特性を電気等価回路で再現した声道アナログ型合成器が提案された。1970 年代には電電公社電気通信研究所が線形予測分析(LPC)合成系を提案し、信号処理を大幅に軽量にすることができた。これを用いて 1978 年に TI(テキサス・インスツルメンツ)がゲーム機 Speak&Spell で一定数のメッセージを音声出力することに成功した。

任意のテキストから音声を合成する研究は、MIT の Klatt が職人芸的に韻律規則や音韻接続規則を記述しすることによって初めて実現させた。これをベースに米国 DEC が 1983 年に DECTalk を製品化した。その後、コンピュータの処理能力の向上とともに音韻波形を編集加工することが可能となり、波形編集方式の研究がされ、合成音声の明瞭度が向上した。

1990 年代には、さらに音のつながりが滑らかなで自然な合成音声を目指し、韻律規則や音韻接続規則を実際のデータから導出する研究が進んだ。1990 年代後半になると、音声認識では標準的な手法である HMM 法により音韻のセグメンテーションを行い、合成のための基本音韻データを自動的に作成することが可能となった。これにより、合成データの作成が大幅に自動化され、あるの人の基本音声データを収集すれば、その人の音質の音声合成システムが比較的容易に作成できるようになった。

(2)現状と課題

現在のPCソフトウェアによる音声合成はイントネーション、明瞭性ともに通常の人の声に非常に近く、それほど違和感なく使えるレベルに達している。今後は、朗読調、対話調など様々なスタイルの合成や感情を付与することなどが課題である。

3.2.3 自然言語処理

(1)発展の歴史

自然言語処理とは、人が通常のコミュニケーションに用いている言語を、コンピュータにより理解したり生成したりする技術であり、人がコンピュータと直接コミュニケーションする場合の基本技術である。自然言語処理の研究は、1950 年代のコンピュータに翻訳をさせる試み(機械翻訳)の研究から始まった。米国において露英翻訳の研究が始まり、日本でも九州大学、電総研で開始された。

米国では、1966 年に「機械翻訳の品質が悪く、コンピュータパワー不足から翻訳品質改善は当面困難であるので基礎的な研究を推奨する」との ALPAC レポート(米国の NSF が組織した自動言語処理に関する委員会レポート)が出され、研究の中心が言語学の基礎研究へシフトして、機械翻訳の研究は停滞した。

1970 年代後半には、翻訳ニーズの強い欧州やカナダで機械翻訳の研究が盛んになり、言語的に近い言葉間の翻訳方式として、対象言語間の単語の訳語を対応付けるトランスファー方式が開発された。これに基づき、例えば、カナダでは 1976 年に天気予報の英仏翻訳が実用化された。

一方、日本では、日本語と英語の言語距離が大きいため構文まで解析する、構文トランスファー方式が研究された。1980 年代前半に国のプロジェクトとして京大を中心に科学技術論文抄録の日英・英日翻訳が研究され、1986 年には日本の電機メーカーから汎用コンピュータベースの日英・英日翻訳システムが製品化された。この翻訳システムの品質はそのまま翻訳文として使えるレベルにはなかったが、人間による翻訳の効率アップのための有用な支援システムとなった。

1990 年代に入り、それまでの人手による文法・辞書作りの限界に対処するため、大量の日英対訳文データから文法や辞書を作成する技術の開発が進展した。コーパスベース翻訳と呼ばれ、翻訳品質が向上した。また、コンピュータ性能の向上から 1990 年代前半は WS ベース、1990 年代後半には PC ベースの翻訳ソフトウェアが製品化されるようになった。

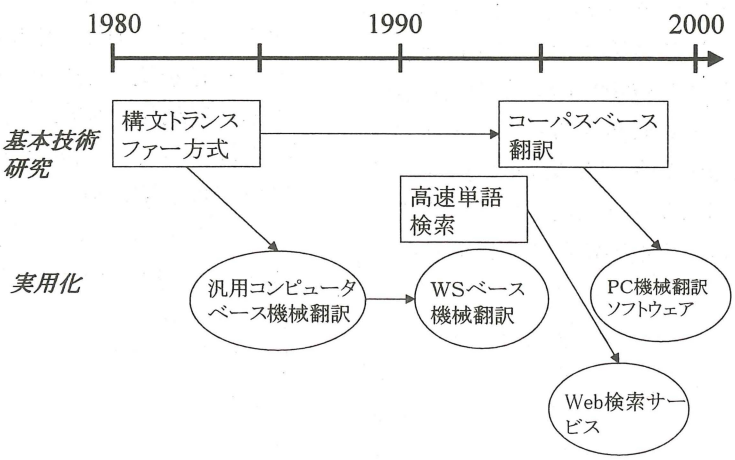
一方、自然言語処理技術の応用として、キーワードによる検索技術が 1990 年代に開発され、文書の自動分類や検索が可能となった。この技術には、機械翻訳研究で確立された単語の切り出し品詞の割り当て技術(形態素解析)が使われている。また、Web サイトの検索には Web 検索サービスが欠かせないものとなっているが、この実現には、並列コンピュータによる高速単語サーチ技術と形態素解析が組み合わされている。

(2)現状と課題

現在の英日機械翻訳システムの性能は、アジア太平洋機械翻訳協会(AAMT)技術動向調査委員会の報告によれば、TOEIC700 点程度(英語力中級クラス)以下の人にとっては、読解力の向上に役立つレベルに達している。機械翻訳の性能を更に向上させるためには、現在、解析できていない言葉の意味的つながり、文脈的つながりの研究を進める必要がある。

また、Web 上の情報は日々増大しており、この膨大な情報源からその人にとって有用な情報を取り出し、整理することが期待されている。ここでは、単なるキーワード検索を超えて、意味や概念まで利用した検索、要約が課題となる。

図表 2 自然言語処理研究と実用化の流れ



3.3 次世代ヒューマンインタフェース技術推進への課題

前章で述べたように、音声認識、自然言語処理ともに、人間の機能をコンピュータで実現させようとする試みであり、長い継続的な基礎研究の中からブレークスルーが生まれ、それを基に応用展開、製品化が実現してきた。しかし、まだその実現レベルは満足いくものではない。より人の機能に近づいた次世代ヒューマンインタフェースを実現するための研究の方向、研究環境、研究マネジメントの課題を述べる。

3.3.1 研究課題

音声認識や自然言語処理の性能は、1990 年代に統計的手法により大きく進展したが、現在大きな壁に直面しているように思われる。現状の音韻認識率は人間より劣っており、また、韻律情報は使われていない。意味処理、文脈処理、対話からの意図抽出・状況判断ができていない。

統計データからアプローチするというモデルに、積極的に音響学、言語学の知見の取り込み拡大し、現在の枠を打破する新しいモデルを創出することが望まれる。図表 4 に日本における代表的な音声認識研究プロジェクトの中にも、基礎的要素技術の先駆け的研究があり、その成果に期待したい。

長期的な視点で見れば、認識とか言語理解などは、人の脳で行われている認知や学習と深く拘わっている。20世紀はディジタル情報処理技術が大きく進展したが、認知や学習のメカニズムは解明できていない。人の脳に学ぶ必要がでてこよう。21世紀における大きな課題の一つであろう。

3.3.2 共通基盤データベースの整備

音声認識や自然言語研究の基盤としてデータ収集が重要であるが、その収集は膨大な労力を必要とし、個々の研究機関ですべてを開発するのは困難である。すなわち、膨大なデータを収集する労力に加え、さらに収集したデータに正しい解析結果が付与されているかを、人手でチェックしなければならず、この作業に多くの労力を必要とし、一研究機関にて開発することは困難である。

言語知識データベースを収集・蓄積し、それを会員間で共有する目的で公的支援をベースにした会員制コンソーシアムとして、米国では 1992 年に LDC(Linguistic Data Consortium)が、欧州では 1995 年に ELRA(European Language Resource Association)がそれぞれ設立され継続的に活動している。これらの組織では、専門のスタッフを抱え、データの収集、保守、配布を行っている。研究用には廉価で提供し、また高価となるが商用使用も許可している。

一方、日本でも共通データベース作成の試みがなされているが、単発的で継続的な活動につながっていない。プロジェクト終了とともにそのプロジェクトで構築したデータベースが消えてしまうことが多い。プロジェクト終了後に維持管理、高度化のための資金が得にくいためである。

1999 年には、日本でも欧米に見習って共通データベース収集・維持・拡張のための組織である言語資源共有機構(GSK)が発足したが、資金不足で実質的活動は始まっていない。研究の基盤整備からは直接研究成果はでないため、研究資金が得にくいのが現状である。

さらに、共通のデータベースは研究の基盤であるだけでなく、このデータベース上で研究されているシステムの性能評価も公平に行うことが可能であり、研究機関同士の公平な競争を促進する上でも共通データベースの整備が望まれる。

3.3.3 研究プロジェクトの日米の相違

研究開発成果を実用に結びつけるスピードの日米格差が拡大していると言われているが、音声認識技術の実用化においても、それが見られるケースがある。

米国における音声認識の研究開発に DARPA が果たした役割は大きい。図表 3 に示すように、音声認識の応用イメージを明確に示してプロジェクトを実施した。このプロジェクトでは同じ目標実現に向け複数の研究機関に対して資金を提供し、競争させることにより目標の早期実現を図った。具体的には、プロジェクト開始とともに評価のためのデータベースを構築し、メンバー間で共有した。研究機関は大学が中心であったが、各要素技術を統合し応用シーンをデモンストレーションする研究試作システムを作成し、その性能を競った。その後、プロジェクトの成果である知的財産権(IPR)は研究実施機関に移管され、その研究機関の意志のみでベンチャーなど企業へ技術移転が進められた。具体例を、DARPA が 1990 年から 1994 年まで実施した航空旅行情報に関する音声対話方式の研究に見ることができる。1995 年には研究実施機関であった MIT, CMU(カーネギーメロン大), SRI の研究者がベンチャー2社をボストンとシリコンバレーに設立した。電話によ

図表 3 米国 DARPA プロジェクトと成果の実用化

年代	プロジェクト	プロジェクトの目標	成果の実用化
1980年代後半 1990年代前半	ディクテーションプロジェクト 小規模タスク読み上げ(1987-1990) 新聞読み上げWSJ(1991-1996)	新聞の読み上げ音声のディクテーション	PCソフトウェアDictationの発売(IBMなどから)
1990年代後半	Q&A対話プロジェクト 航空旅行情報検索ATIS(1990-1994) Web情報検索TIDES(1998-2000) Communicator(1999-2003)	音声対話による航空券予約、天気予報問い合わせ	コールセンターでの音声対話システム(米国ベンチャー企業2社から)
2000年代前半	Human Centered System音声エージェントプロジェクト(2002-2007)	音声電子秘書エージェントの構築	

る各種予約・問い合わせシステムを構築しコールセンターサービスに音声認識を導入して、省力化、24時間サービス化を図った。現在米国では、電話による航空券予約サービスに加え、各種の予約・情報提供サービスにコールセンター音声自動化システムが使われ始めており、この2社がほとんどのシェアを押さえるまでに成長している。

一方、日本においては、図表 4 に示す音声認識の代表的な国のプロジェクトが実施されている。この中でATR の「音声翻訳研究プロジェクト」と最近の「韻律に着目した音声言語処理の高度化プロジェクト」を除くと、米国のDARPAプロジェクトとほぼ同様に朗読音声、対話音声の性能向上を目標としたプロジェクトである。「音声言語による人間・機械対話システム」では研究試作システムまで作成し評価を行っているが、それ以外のプロジェクトでは、研究の目標を各要素技術に分解し、各大学で分担して進める方式であった。要素技術の向上には貢献したが、プロジェクトでは要素技術を統合した評価システムまでは作っていないため、実際全体システムとしてどの程度の性能を持つかは不透

図表 4 日本における代表的音声認識プロジェクト

時期	プロジェクト名	内容	投資形態
1986～ 1989年度	音声言語によるマン・マシン・インターフェースの高度化	朗読音声を対象とした音声認識の研究	文科省科研費重点領域研究
1993～ 1995年度	音声・言語・概念の統合的処理による対話理解と生成に関する研究	音声・言語統合処理による対話理解の研究	文科省科研費重点領域研究
1993～ 1999年度	音声翻訳研究プロジェクト	ホテル予約の会話音声の日英双方向自動通訳	ATR(旧エイ・ティ・アール音声翻訳通信研究所)
1996～ 2000年度	音声言語による人間－機械対話システム	インターネット上での文献検索の音声対話システムの実現	日本学術振興会未来開拓学術研究推進事業
1997～ 1999年度	日本語ディクテーション基本ソフトウェア	日本語連続音声認識基本ソフトウェアの開発	IPA(情報処理振興事業協会)プロジェクト
1999～ 2003年度	話し言葉工学の構築	話し言葉から意味・内容・話しての意図などを抽出する技術の基盤の確立	科学技術振興調整費開放的融合研究推進制度
1999～ 2003年度	多元音響情報の統合的理解	空間音響、音声分析合成、音声認識、対話、音情報認知の研究	文科省COE形成プログラム
2000～ 2002年度	擬人化音声対話エージェント基本ソフトウェアの開発	機械が人間のように対話する技術の基本ソフトウェアの開発	IPA(情報処理振興事業協会)プロジェクト
2000～ 2003年度	韻律に着目した音声言語処理の高度化	基礎から応用にわたる韻律研究の統合	文科省科研費特定領域B

明であった。一方、産業界は学会での知見をもとに独自に応用開発し、製品化してきた。産学連携は活発とは言えず、米国のような大学からベンチャーへの技術移転も起きていない。言い換えれば、日本の今までの国のプロジェクトでは産学連携を意識したものではなく、基礎研究のレベルアップが中心であった。

日本音響学会の「音声言語関連大型プロジェクトの現状と将来」のパネル討論[1]で国のプロジェクトに拘わる問題点として、まったく斬新な、リスクの多いアプローチがとりにくいこと、プロジェクト終了後のアフターケアのないことが上げられている。すなわち、日本の今後のプロジェクト制度においては、「成功を当然とする文化」から「失敗を恐れず真剣に競争する文化」に変えること、プロジェクトの成果を次のプロジェクトに有効に継承していくことなどが対処すべき課題である。

また、ヒヤリングした大学教授からは、情報分野ではポスドクなど大学の研究スタッフは非常に少ないため、大きなシステムを作成するパワーはなく、細かな要素技術研究に走りがちであるとの指摘があった。

米国の大学では、ベースの研究資金は少なく、大学外部から研究資金を獲得する必要がある。国家プロジェクトは重要な研究資金の一つであり、その獲得競争のため基礎研究といえどもその目標を明確にしている。また、DARPA音声認識プロジェクトの例で言えば、公平な競争による研究加速が見られたと言われている。研究成果の実用化が迅速かつスムーズに行われていることも、国全体として、技術開発効率アップにつながっている。

3.4 おわりに

次世代ヒューマンインタフェース技術には、情報機器が広まる中でDigital Divideの解消、初心者を含む利用者の使い勝手の向上を実現することが望まれている。また、あらゆる情報がネットワークにつながり、そのアクセスもモバイル機器を介して行う時代が目前にあり、いつでも、どこでも、情報へアクセスすることが実現しつつある。情報機器との自然な対話によって欲しい情報を簡単に入手することが究極の目標であろう。

これを実現するためには、多くのハードルがあるが、一つは、基礎研究の活性化が求められている。現在、性能向上はある面では頭打ちになっており、これを打破する新しいモデルへの挑戦が必要であろう。

世界的に企業が基礎研究を大学に依存する傾向が強まっており、大学の基礎研究への期待が大きい。リスクは大きいが斬新なアイデアで研究プロジェクトが進められるように、真剣な競争と評価、さらに失敗を許容し再チャレンジができる文化の醸成が必要であろう。

さらに、創出されたブレークスルー技術を迅速に実用化へ向けて努力を集中させる必要があるが、米国に比して弱い産学連携、技術移管を強化することも課題である。

用語説明

①ゼロ交差数

波形がゼロを交差する数であり、波形のエネルギー量の近似値が得られる。波形の特徴量をハードウェアにて比較的容易に求められる方式として使用された。

②DPマッチング法

音声中の各音韻の発声時間は発声するごとに伸縮する。この伸縮を正規化した上で最も似ているパターンを求める方式として動的計画法(Dynamic Programming)を用いたもの。

③コーパスベース翻訳

コーパスとは実際の膨大な例文とその実翻訳文のデータベースを示す。コーパスベースの機械翻訳では、コーパスから抽出した文法規則や翻訳の用例をそのまま適用する。従来の経験則ベースの機械翻訳に比較して翻訳品質が向上している。

参考文献

1) 藤崎, "パネル音声言語関連大型プロジェクトの現状と将来のために",情報処理学会音声言語情報処理研究会 29-40 (1999)