T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	究極の音声認識・理解を目指して
Title(English)	
著者(和文)	古井 貞熙, 有木 康雄, 岡 隆一, 鹿野清宏, 中川 聖一, 牧野 正三
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会シンポジウム：パターン認識・理解の新たな展開にむけて，，，
Citation(English)	，，，
発行日 / Pub. date	1993, 9
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1993 Institute of Electronics, Information and Communication Engineers.

Powered by T2R2 (Science Tokyo Research Repository)

究極の音声認識・理解を目指して

Towards the Ultimate Speech Recognition and Understanding System

音声分科会主査：古井貞熙[1)]
委員：有木康雄[2)]、岡　隆一[3)]、鹿野清宏[1)]、中川聖一[4)]、牧野正三[5)]

1) NTTヒューマンインタフェース研究所　2)龍谷大学理工学部電子情報学科
3)電子技術総合研究所　4)豊橋技術科学大学工学部情報工学系
5)東北大学応用情報学研究センター

あらまし　これから目指す究極の音声認識・理解システムの姿は如何なるものであるべきかについて検討し、そのようなシステムを実現するために必要な本質的課題にどのようなものがあるかについて述べる。それらの課題を、現在の音声認識・理解研究のアプローチの発展として考えられるものと、比較的新しい発想に基づく課題に分類した。最後に具体的な研究課題として、音声を特徴づけているものは何か、音声現象と識別学習、ディクテーションマシン、実時間音声会話アミューズメント・システムの構築問題、音声認識と自然言語処理との融合、話者認識、音声言語の識別、の７種類の課題について、やや詳しく述べる。

1　まえがき

音声認識・理解の研究は長い歴史を有し、近年は、不特定話者が発声した大語彙の連続音声の認識を目的とした比較的基礎的な研究から、具体的な実用システムの実現を目的とした研究開発まで、広い範囲にわたった活動が活発に行なわれている[1][2]。表１に、音声認識・理解システムの、今後の実用化レベルの発展の予想を示す。これはRabiner, Juangのオリジナル[3]に古井が手を加えたもの[4]である。これらのシステムの主な用途としては、データベース（情報）検索・案内・操作、予約、注文、ディクテーション、音声ワープロ、電子秘書、ロボット、自動通訳電話、障害者の補助などが考えられている。

音声認識・理解システムの究極の姿をどのように考えるかによって、あるいはどの程度の時間のレンジで考えるかによって、今後研究すべき課題が当然変わってくる。ここでは、長期的な視野から、限りなく人間の能力に近い能力を有し、なおかつ幾つかの点で人間の能力を越えることができる音声認識・理解システムを究極の姿と考える。人間の能力を越える方法としては、より早い、より正確、より賢い、より安い、より気楽に使える、など種々の側面が考えられる。そこで、そのようなシステムを実現するために重要と考えられる本質的課題にどのようなものがあるかを考え、それらを、現在の音声認識・理解の研究のアプローチの発展として考えられるものと、今後の比較的新しい課題に分類する。前者については２章で、後者については３章で説明する。４章では、具体的な課題を個別に取り上げて解説する。

表1 音 声 認 識 ・ 理 解 の 今 後 の 展 開

年	1990	1995	2000	2000+
認識方法	孤立／連続単語 単語単位モデル ワードスポッティング 有限状態文法 制約の大きいタスク	連続音声 単語より小さい認識単位 統計的言語モデル	連続音声 単語より小さい認識単位 自然言語寄り言語モデル タスク依存意味モデル	連続音声 対話音声向き構文・ 意味モデル 適応化・学習
語　彙	10 - 30	100 - 1,000	5,000 - 20,000	無制限
プロセッサ	2 ～ 4 DSP (25Mips / Chip)	4 ～ 10 DSP (50Mips / Chip)	5 ～ 50 DSP (200Mips / Chip)	20 ～ 60 DSP (1,000Mips / Chip)
応　用	音声ダイヤル 簡単な情報案内 （バンキングサービス） 簡単な予約サービス	自動オペレータサービス 音声コマンド 商品の注文・予約 情報案内	ディクテーション 電子秘書 データベースアクセス 電話番号案内	コンピュータとの音声対話 自動通訳電話

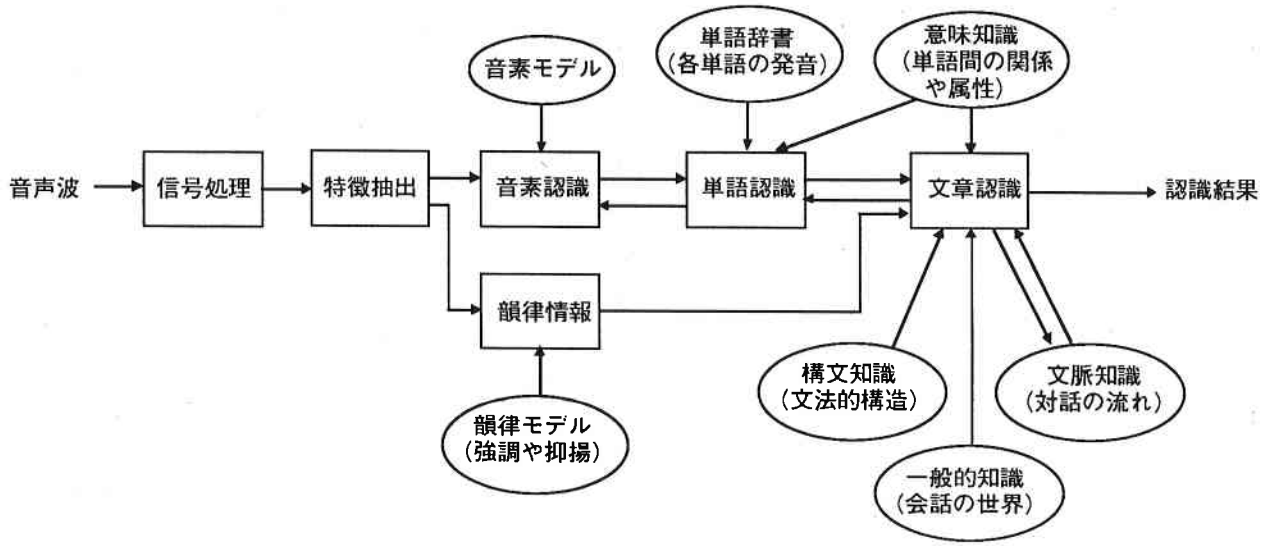

図1　音声認識・理解システムの典型的構成

2　現在の研究のアプローチの発展としての課題

音声認識・理解システムの典型的な構成を、図1に示す。

現在の音声認識・理解の研究のアプローチの発展として考えられる重要課題を、この音声認識・理解のプロセスの順序にほぼ対応させて、(1) 信号処理とアルゴリズム、(2) モデリング、(3) 特徴空間、(4) 記号化の単位、(5) 学習方式、(6) 認識手法、(7) 言語処理、(8) 音声データベース、に分類し、さらに大きく、比較的長期的な課題と、当面成功の見込める課題に分類して、表2に示す。

音声認識・理解の方法を評価したり、新しい方法を考える上で、非常に重要な視点は、種々の要因による音声の変動に対して耐性の高い方法でなければならないということである。そのような方法でなければ、本質的ではないということもできる。表3に、音声の種々の変動要因を分類して示す[5]。これらの変動に対する耐性を持たせる最も基本的、かつ重要な方法は、システムに、自動的に変動に適応できるしくみ（適応性）を持たせることである。

表3 音声の主な変動要因

環境	音声と相関のある雑音： 反響（残響）、反射
	音声と相関のない雑音： 付加雑音（定常、非定常）
話者	個人性：　個人差、方言、性別、年令
	発声方法：呼吸音、不要音（口唇音など） アクセント、イントネーション ストレス、Lombard効果 発声速度 発声レベル（声の大きさ） ピッチ（声の高さ） 丁寧さ（協力的、非協力的）
入力機器系の電気的雑音／歪	マイクロホン（送話器） マイクロホンと話者の相対関係（距離など） フィルタ 伝送系（歪、雑音、エコーなどを含む） 録音系

3　今後新しく研究すべき課題（今後挑戦すべき本質的課題）

3.1　対話音声の認識・理解

音声認識・理解システムの応用形態は、対話システムとディクテーションに分類できる。前者には、電子秘書、データベース（情報）検索および操作、などが含まれ、究極の応用として、自動通訳電話が考えられる。これまでの音声認識・理解の対象は、書かれた文章を読み上げた音声を対象とするものが多く、文法的、あるいは意味的にほぼ正確なものを対象としていたが、対話音声のようなill-formedな世界の処理を可能とするためには、従来になかった新しい発想が必要である。すなわち、逸脱した文法、発音、不要語、未知語などを取り扱える枠組みを構築することが必要であり、その具体的方法として、キーワード駆動型音声理解システム、すなわち、構文よりむしろキーワードの意味をベースに音声理解を進めるシステムが考えられる。意味を中心とした文節程度のローカルな構文制約に、比較的ゆるい文レベルでの構文を組み合わせる方法も考えられる。また、会話音声でのキーワードのスポッティング手法の適用拡大として、スポッティングする単位を文レベルまで大きく拡大して認識する方法も考えられる。

対話音声の認識・理解のためには、韻律情報の利用も必要性と重要性が増してくると考えられる。将来的には、感情などの心理的状態の音響的表現など、感性情報の取扱いも必要となってくるであろう。

表2　現在のアプローチの発展として考えられる音声認識・理解の重要課題

	長期的課題	当面の課題（当面成功の見込める課題）
信号処理とアルゴリズム	・頑強で適応性のある認識アルゴリズム 　残響、周囲雑音、マイク、回線雑音の発生機構を考慮した統計的信号処理 　少量の音声を基に如何に新しい話者や環境の変化に適応するか ・信号処理の精度の音声認識における充分性の理論的解析 ・インパルスレスポンス長より短いパルス間隔で駆動された信号の分析法（女声の分析）	・位相情報やピッチ情報を利用したスペクトル推定法 ・ワードスポッティング技術の利用（耐雑音性） 　ワードスポッティングで不要語や、雑音を取り除き、音声のみを認識する ・長時間（数十～数百ms）をベースとした信号の分析法
モデリング	・調音結合のモデル化と正規化 ・音声生成や音声知覚の機能に基づいたモデル化 　スペクトルの奥に潜むメカニズムに基づいて音声現象をモデル化する ・確率の概念の新しい展開とＨＭＭ以外のモデルの探索 ・話者の特性を表す動的統計モデル ・視覚のearly visionのようなearly speech(auditory)モデル ・韻律情報の統計的モデリングと統計的音声認識への組み込み	・不特定話者に対して頑強な認識モデル ・情報理論、確率過程をベースとした音声分析・認識・言語モデルの拡充 　モデルの拡張、モデルの同定・学習・適応・汎化など ・ニューラルネットワークの音声認識・理解システムへの組み込み 　ニューラルネットを時系列を扱えるよう拡張し、認識率をあげる ・単純なアーキテクチャとの適合問題 　透明性のある単純なアーキテクチャを持つようなデータ構造とアルゴリズムの開発
特徴空間	・最適な特徴パラメータは、言語・音韻に独立か ・自由ベクトルを表現できるパターン空間 　座標原点に影響されない自由ベクトルによる特徴パラメータ表現 ・特徴ベクトルの物理空間から感覚空間へのマッピング 　物理的量の特徴ベクトルを感覚量で表現した空間へ写像し単語を認識する ・特徴パラメータの評価法 　抽出した特徴パラメータが認識に対して十分かどうかの評価	・動的特徴パラメータの抽出と表現法 　変化に含まれている特徴の安定な抽出方法 　フレームレベルの動的特徴か音韻レベルの動的特徴か 　話者に不変な動的特徴か話者に依存する動的特徴か 　静的特徴または従来特徴と動的特徴の関係 　音声の発声機構と動的特徴との関連
記号化の単位	・初期記号化の要素単位の決定と初期記号の統合化 フレーム単位の特徴ベクトルを用いる場合、それから最初に記号化される単位を何にするかの決定問題	・音声認識単位の粒度を状況に応じて動的に変える技術 ・コンテキスト・環境ごとの標準パターンの作成か、コンテキストの変動も記述・包含した標準パターン（たとえば、混合分布）の作成か
学習方式	・認識手法の自動学習 　位相情報の認識への利用といったことが統計的手法で学習可能か ・モデルの構造的学習と統計的学習の両立 　データからモデルの構造と統計的パラメータをともに学習する枠組み	・少数サンプルからの母集団あるいは標準パターンの推定 　標準パターンから実際のサンプルを生成する共通の伝達関数を仮定する ・学習内容を評価する各種の基準リストとその形式化 ・誤差最小化（確率最大化）基準がよいか、誤り数最小化基準がよいか（特に耐性の観点から）
認識手法	・構文的手法と統計的手法の統合 　統計的手法と構文的手法をどのレベルでどのように組み込むか ・部分パターンからの全体の推定と統合 　数十ミリ秒の部分パターンから音素・音節を推定し全体を統合する 　理解に不必要なところは気を抜きながら必要なところのみ聞き取るモデル ・局所的処理と大局的処理の統合 　情報量の多い局所的な情報をもとに大局的な処理を制御し認識する ・認識系全体の制御問題 　ネットワークモデルと階層モデルの融合のような新たな制御方式の定式化 　モジュールの相補性によるロバストネス解析 　オブジェクトオリエンテッドなモデルによるシステム構成 　フレーム同期性制御の新しい展開 ・韻律情報の構文意味解析への利用	・未知語の処理 　未知語を未知語として認識し、結果を音節単位で出力する ・棄却率の向上 　棄却率を少なくし、認識性能を高める、あいまいな結果を正しく棄却する 　棄却した単語の自動学習 ・認識尺度として尤度がよいか事後確率がよいか（ML推定かMAP推定か） ・欠陥パターンからの特徴抽出と比較 　雑音などにより破損した欠陥パターンと標準パターンとのマッチング ・ヒューマンインタフェースとしての評価と技術開発 　ヒューマンインタフェースの観点から技術を正しく評価し、それに基づいてバランスよく新しい技術を開拓していく ・音声入力を必要・不必要とするタスクの特徴づけ ・音声認識結果を画像で表現するとき、画像による音声認識結果の補完性 ・音素コンテキストモデルの連結のための音声生成モデル
言語処理	・方言、性別、年令等による言い回しの違いへの対処 ・発話行為、感性情報の処理	・統計的言語モデルの適応化 ・対話におけるポーズや間投詞の役割の分析 ・ill-formedな発話文の処理 ・概念等の共起関係や連想の利用
音声データベース	・音声データベースの作成問題 　どの様な認識課題に、どの様な音声データを、どの程度集めればよいかを決定する方法を明らかにする ・データベースからの知識の自動獲得	・不特定話者連続音声データベースの構築

3.2 学習方式

前章でも述べたように、種々の変動に対する耐性を持たせるために、認識モデルが動的に変更、適応化できる機能をもつようにすることが必要である。このためには、環境モデル、ユーザモデル、話者モデルなどを認識過程で学習することができなければならない。

さらに一歩進めて、音韻、音節、単語、文法、概念などの獲得を、音声および言語データからの種々のレベルでの学習により、自動的にできるようにする研究も、長期的には重要である。これが可能になれば、単語の構文情報や意味情報、単語間の接続関係などを自動的に獲得できるようになり、さらに話題の変化などによるその変動に、自動的に追随できるようになると期待される。

究極の音声認識・理解システムは、語彙、構文、意味、タスクなどに制限のないシステムであろう。これも、広い意味での自己学習型システムによって、可能になると考えられる。

3.3 音声処理と言語処理

音声処理と言語処理はこれまで分離して研究されていたが、今後はこれら二つの処理を融合した統合モデルの研究を進める必要がある。特に今後、対話音声を対象とするためには、話言葉としての音声処理に適した言語知識と言語処理を研究し、構文・意味に関する言語的知識を状況に応じて音声的（音響的）知識に並列的に組み合わせることを可能とする必要がある。

さらに、音声から概念への写像と、概念空間での連想による音声認識のアプローチにより、連続音声中の単語を物理的観測空間のみでなく、概念空間で意味的に認識するという発想も必要であろう。また、周囲の雑音は、これまで排除あるいは除去することだけを研究してきたが、音声だけを認識するのではなく、環境音の認識も含めて状況を理解すること、すなわち、車の音などの雑音を積極的に認識して状況理解に役立てることも必要となるであろう。

3.4 音声認識・理解システムのアーキテクチャ

音声特徴抽出、認識、理解、および対話処理全体を統合するアーキテクチャのあり方を研究することも重要である。今後の高度な音声認識・理解システムは、現在に比べて一層複雑かつ大規模になることは明らかであり、並列処理プロセッサやプログラムをいかにうまく使いこなすかということも、実際のシステムを動作可能とする上で重要なファクターとなるであろう。種々の処理の統合に要する透明性と単純性によって、アルゴリズムを評価することも、長期的には必要になってくると思われる。

3.5 音声認識・理解システムの評価法

音声認識・理解システムの評価法は、次の二つに分類できる。

(1) タスクの評価法：対象とするタスクの複雑さと難しさを評価する
(2) 技術の評価法：技術のレベルを評価する

認識・理解性能は、システムが対象とするタスクの難しさによって、当然変化するので、その性能をタスクの難しさで適切に補正しないと、互いに善し悪しを比較することができない。このため、タスクの評価法は極めて重要である。現在広く用いられている尺度に、単語、音節、文字、音韻などを単位とするパープレキシティがある。これは情報量的な尺度で、単純な認識システムの評価には便利な尺度であるが、意味を考慮するのは難しい。文章の意味を理解するという意味での音声理解システムの難しさの評価ができる尺度を構築するのが、今後の課題である。

技術の評価法に関しては、音声認識・理解システムは人間が使うものであるから、その善し悪しの評価法は、ユーザの主観を反映したものでなければならない。すなわち、ヒューマン・マシン・インタフェースの改善にどれだけ役に立つか、ユーザにとってどれだけ使いやすいか、という視点から評価できなければならない。その尺度としては、次のような項目が考えられる。

(1) 認識性能
(2) マルチメディア環境の中で、種々のメディアの役割分担が適切であるか
(3) 誤り傾向などがユーザにとって自然であるか
(4) 認識誤りの修正が容易であるか
(5) 修正した誤りを繰り返すことはないか

認識誤りが少ないに越したことはないが、単純に少なければ良いというものではない。誤りには、置換、挿入、脱落、など種々の現われ方があり、それらの間にはトレードオフがある。それらをどう評価するか、すなわち、それらへのペナルティーの配分は、実際のシステムの使用状態を十分考慮して行なう必要がある。

今後のヒューマン・マシン・インタフェースとしては、従来の電話機をそのまま用いたサービス以外では、すべての入出力を音声で行なう必要はない。入力に関しては、音声認識のみにこだわらず、文字、アイコンなど、種々のメディアを適切に組み合わせて用いることが必要である。システムの評価も、このような観点から行なうことが必要である。

仮に認識誤りがある程度あっても、それがユーザにとって自然で、人間の感覚と合っていれば、受け入れられやすい。さらに、その誤りをユーザが容易に修正でき、その後に、ユーザが同じ言葉を改めて発声したときには、同じ誤りを繰り返さない学習機能を持たせることが必要である。

3.6 人間に学ぶ

3.6.1 人間に学ぶ必要性

音声認識・理解の研究は（パターン認識・理解に共通であるが）、あらゆる人間が制約なく自然に発声した音声を対象とするという意味で、自然科学的側面を有している。また、言うまでもなく、音声

は人間が発声器官を動かして生成するものであるから、それらの器官の運動上の特徴や制約を反映している。さらに、一般にそれは誰かに聞かれることを前提として発声され、自らの聴覚を通じたフィードバックに基づいた発声器官の制御に基づいているので、人間の聴覚器官の働きとも密接な関連を有している。

これまで、音声生成や音声知覚のメカニズムに関する研究の成果が、音声認識・理解の研究の進展に直接役に立ったことは少ないが、今後、ユーザにとって、より自然で、耐性の高いシステムを実現するためには、人間の音声生成および知覚機能の解析を進め、その原理を取り入れたシステムを構築する必要性が増してくると思われる。

3.6.2 音声生成メカニズム

音声生成における実際の音源は、現在広く用いられているモデルにおけるような、単純なパルスや白色雑音ではなく、しかも、調音フィルタと線形分離可能であるとは必ずしも言えない。このため、音源、および、音源と調音フィルタの相互作用に関する本格的な研究が不可欠である。

音素や音節が連続して発声されるとき、舌、顎、口唇のような調音器官は、並列、非同期で、しかもシステマティックに動く。現在の音声分析技術では、音声信号を短時間スペクトルの系列として一面的にしか見ないが、今後、音声生成のメカニズムに基づいて、隠れた複数の要因に分解して取り扱うことが、重要になるであろう。すなわち、調音器官の運動規則の合理的記述ができれば、音響レベルでは記述の難しい調音結合現象を、より簡明に記述し、その規則を内蔵したモデルを構築できる可能性がある。それらの研究が進めば、音声の個人差、調音結合、区分化（セグメンテーション）の問題の解決の糸口がつかめると期待される。

3.6.3 音声知覚メカニズム

人間における音声の生成と知覚のメカニズムは、密接な関連を有している。今後の音声認識・理解の大きな飛躍を成し遂げるための新しい指導原理を得るために、人間の聴知覚に関する心理・生理学的研究が必須になってくると予想される。

人間の聴覚の、心理学的および生理学的研究から、聴覚器官は音の変化に対して極めて敏感であることが分かっている。音声知覚においても、音声の動的情報が重要な役割を果たしていることが、いくつかの実験で確認されている。音声の動的情報が知覚される時間窓の長さは、数msのオーダから数秒のオーダまで、階層的構造をもっている。これらの階層は、音素、音節、韻律などの種々の音声現象の知覚に対応している。また、人間の聴覚器官には、スペクトルの動きから、その変化の目標を知覚する機能があることも分かっている。これらの観点からの研究により、調音結合現象の把握などへの一歩が踏み出せる可能性がある。

人間の音声知覚においては、言うまでもなく、言語的知識が重要な役割を果たしている。自然言語処理の研究は、これまで人工知能研究の中心テーマとして、知識の表現、蓄積、探索、述語論理、推論などという形で、精力的に研究されてきた。今後は、人間における自然言語処理メカニズムの研究を進め、その原理を利用した音声理解アルゴリズムを構築することが重要になってくると考えられる。

（以上は委員全員による討論をまとめたものである）

4 今後挑戦すべき個別課題の例

4.1 音声を特徴づけているものは何か（音声と雑音の分離）

4.1.1 背景

(1) 音声と雑音の分離の必要性

商用の音声認識装置が実際に用いられる場面は、自動車電話の電話番号入力、電車の券売機、工場でのデータ入力、オフィスでの音声ワープロなどであり、実環境での雑音が必然的に混入する状況にある。工場やオフィスなどでは接話マイクを使用して雑音の混入を防止することも可能であるが、自動車電話や券売機、特に不特定の人が頻繁に使用する状況においては、接話マイクは不向きである。

商用の音声認識装置を、実環境において接話マイクのない状態で使用すると、認識率は大きく低下する。実環境で精度よく音声認識装置を稼働させるためには、雑音の混入を未然に防ぐ工夫と、音声信号を破壊することなく、混入した雑音を除去する方法の開発が必要である

また、電話のような音声通信においても、送信者側の環境から生じる雑音が音声に重畳するので、音声の明瞭度が低下するといった問題がある。このため、音声に重畳した雑音を除去して、明瞭な音声として送信または受信する必要がある。将来的には、実環境において生じている雑音を除去するのではなく、逆に認識して、発話者の環境を理解した上で音声認識を行なうことも必要となろう。

(2) 現在の到達点

音声認識の研究は、人間の音声を機械にいかに認識させるかを中心課題としている。従って、人間の音声、特に言語音にはどの様な特徴があるのか、その特徴を機械でどの様に扱えば音素、音節、単語、文節、文が認識できるのかといったパターン認識の枠組みで研究が進められてきた。通常、商用化されている音声認識装置は、こうした研究の成果をもとに開発されている。この音声認識装置に、雑音処理機能を付加する場合、主として次の２つの方法がある。

(a) 音声の構造を利用しない方法

無音区間において雑音スペクトルを推定しておき、雑音が重畳した音声のスペクトルからこれを差し引いた後（スペクトルサブトラクション法）、雑音のない音声テンプレートとマッチングをとる方法が代表的なものである。また、これとは逆に、音声テンプレートの方に雑音を予め付加しておき、雑音の重畳した入力音声を認識する手法も研究されている。この方法の改良されたものとして、免疫学習法がある。この方法ではまず、雑音の重畳していない学習用音声に、ＳＮ比を徐々に下げながら雑音を重

畳させ、単語スポッティングを行って単語を切り出す。次に切り出した単語を、雑音の重畳した音声テンプレートとして学習するもので、希望するＳＮ比を持つ音声テンプレートが得られるまでこの処理を繰り返している。

(b) 音声の構造を利用する方法

母音は声帯の振動によって作り出される断続した気流、すなわち声帯音源が声道において共鳴を受け、唇から空間へ放射されることによって生成される。雑音が音声に対して加算的であり、互いに無相関であれば、雑音の重畳した母音の自己相関関数は、母音の自己相関関数と雑音の自己相関関数の和として表わされる。母音の自己相関関数には、声帯音源と同じ基本周期が存在するが、雑音の自己相関関数は原点付近に集中し、急速に減少する。従って、雑音の重畳した音声の自己相関関数において、原点付近を外すことにより雑音を除去した音声の自己相関関数を得ることができる。

この方法は、母音では声帯音源に基本周期が存在するという性質を、時間軸上で利用したものである。同じ処理を周波数軸上においても実現することができる。すなわち、声帯音源は、スペクトル上では基本周波数の整数倍の所にピークをもつ線スペクトル構造となる。この線スペクトル構造が母音の情報を担っているので、この部分の周波数成分のみを抽出するコームフィルタを構成することにより、雑音成分を除去するものである。

また、声道における音声の共鳴特性は、線形予測分析によりパラメータとして抽出することができる。この分析手法の中に、重畳する雑音成分を直接表現して定式化し、雑音の重畳していない音声の、共鳴特性のパラメータを推定する方法も研究されている。

以上は、雑音が既に重畳してしまった音声に対する雑音除去方法である。しかし、雑音環境下で音声を収録できる場合には、人間の両耳と同じようにマイクロホンを２本使用し、雑音を相殺して混入を防ぐ方法が開発され成果を納めている。

(3) 現状と目標とのギャップおよび問題点

音声の構造を利用する方法では、雑音が重畳した音声から基本周期や、共鳴特性のパラメータを精度よく推定しなければならないという問題がある。雑音のない場合においては、ある程度確定した方法が存在するが、雑音が重畳した場合に、これらのパラメータを精度よく推定する方法は開発されていない。雑音除去の問題を考える以前に、この方法を開発しておく必要がある。

音声の構造を利用しない方法のうちサブトラクション法では、雑音を除去するとともに、音声の構造も部分的に破壊してしまうため、雑音除去後の音声の明瞭度があまり向上しないといった問題がある。従って、音声認識においても、飛躍的な認識率の向上を期待することはできない。次に雑音を音声テンプレートに重畳させておく方法では、雑音が重畳した入力音声のＳＮ比と、雑音を重畳した音声テンプレートのＳＮ比に差があり過ぎると、認識率が極度に低下するという問題がある。

現在、主として処理されている雑音は、白色性の雑音あるいは、有色性の雑音である。オフィスにおいてはドアをノックする音や回転椅子のきしむ音、靴音といった環境雑音があり、車内においてはクラクションやＢＧＭのような雑音もある。実環境での音声認識装置の適用範囲を広げるためには、この種の環境雑音を除去できる方法の開発が必要である。

4.1.2 主な問題

(1) 雑音が重畳した音声信号から、基本周期とＬＰＣ係数を精度よく抽出する方法を開発する。

(2) 音声認識と音声通信の両方において、雑音除去の効果があるような雑音除去方法を開発する。すなわち、雑音が重畳した音声信号から雑音成分を除去した後においても、音声の明瞭度が高く、かつ音声認識の認識率も高い雑音除去方法を開発する。

(3) 音声認識装置の適用範囲を広げるために、ドアをノックする音、回転椅子のきしむ音、靴音、クラクションやＢＧＭのような環境雑音を、効果的に除去できる方法を開発する。

(4) 二人以上の人が同時にしゃべった場合に、話者一人一人の発話を分離して認識する。

4.1.3 応用例

火事のような非常事態において、「火事」という単語を、多数の話者と騒音の中から抽出して認識する。その単語音声の基本周波数の高さや、声の強さなどを調べることによって、普通の日常会話などで出てきた「火事」という単語には反応せず、非常事態においてのみ反応する音声認識装置が構成できる。

（有木康雄）

4.2 音声現象と識別学習

4.2.1 背景

音声認識・理解システム構築技術のキーポイントは音響処理部の精度の向上にある。言語情報が必要不可欠であるのは言うまでもないが、一般的な音声を対象とした場合、言語処理部の精度向上よりも音響処理部の精度向上の方がしばらくは容易であろうと考えられる。特に隠れマルコフモデル（ＨＭＭ: Hidden Markov Model）に代表されるような確率統計的な認識方法[6]とニューラルネットに代表される識別学習[7]との組み合わせは有望なものと考えられる。そこで、ここでは確率統計的な方法と識別学習を組み合わせた方法を実現する場合の問題点を実際の音声現象と比較して述べる。

(1) 人間の音声知覚

人間の音声知覚では、以下のことが明らかになっている。

(a) 母音の知覚では、定常部のみを聞いた場合よりも子音からの過渡部を含めて音を聞いた方が自然であり、聴取率も格段に向上する。

(b) 子音の知覚では、母音と同じように、母音と子音の間の過渡部、特に子音から母音への過渡部、破裂部などが重要な役割を果たしているが、バズや子音定常部などはあまり重要ではない。またホルマントの移動方向や移動速度が知覚にとって重

要な特徴である。

(c) 連続音声中の音節を正しく知覚するには前後1音節から2音節相当の音声が必要である。

(d) (c)とは逆に、短い過渡部から前後の音素をある程度聞きとることができる。

(e) 個人差や発声速度、雑音に対して柔軟に適応している。適応は非常に速やかに行われる。

以上の知見からわかることは、音声知覚は注目している区間の前後150msから前後300msの流れの中で、30msから50ms程度の物理現象の変化によって行なわれていると言える。このような知覚あるいは認知の機構は文字や図形などにも共通に存在する。例えば、図形における辺とエッジの関係などが音声の定常部と過渡部の関係に類似している。この定常部と過渡部が音声知覚において果たす役割としては、定常部が大分類に寄与し、過渡部が類似音素間の分類に寄与していると考えられる。

(2) 機械による音声認識の現状

現在の音声認識における音素認識の現状は次の通りである。

(a) HMMに代表される音素認識方法では、10msごとのフレームと呼ばれる単位を基本にして標準パターンが構築されている。したがって、標準パターンと入力との尤度計算に対する貢献度はフレーム数に比例する。フレーム数は定常部では長く、過渡部では短い。したがって、現在の方法では定常部間の差異によって識別しているともいえる。

(b) 音声の過渡部の特徴の分散は大きく、定常部の分散は小さい。従って、モデルの尤度に主に貢献しているのは定常部の特徴であり、定常部の変動が識別結果を左右している。

(c) 特徴の方向性が積極的に導入されているが、音素間、音節間を越えたものにはなっていない。

現在の音素認識法は定常部の違いを重視したものであり、人間の音声知覚機構とはかなり異なっている。

(3) 識別学習の現状

人間の知覚に近づけるため、識別学習を利用して定常部の特徴によって大分類を行ない、過渡部の特徴によって類似音素を識別することが試みられている。以下にその代表例を挙げる。

(a) HMMによって音素モデルを構築し、音素対毎に時間構造も考慮して音素モデル間を識別する識別関数を設計し、その統合によって音素を識別する識別学習法が提案されている。

(b) 混合分布型HMMや複数個の標準パターンを用意し、学習ベクトル量子化法によって識別関数を設計する。

これらの方法は識別関数に階層構造を仮定したものであり、優れた発想といえる。しかし、その判別関数は学習サンプルによって固定されたものであり、個人差や調音結合の変動に対処するには柔軟さを欠く。またその学習法は対応する状態間相互の学習に帰結するものが多い。

4.2.2 主な問題

以上の背景説明から明らかになった点をまとめると次のようになる。

(1) 音素や音節の識別は1種類の識別関数で行うことは困難であり、音素の対によって識別特徴、識別関数を変えることが必要である。

(2) 識別関数の設計には時間的な伸縮も考慮できる方法が必要である。

(3) 個人差、調音結合、発声速度、雑音などにも柔軟に対処できることが必要である。

特に(3)の問題点と(1) (2)をどのように融合するかが大きな問題点である。いままでは、個人差や雑音などに対しては多数話者の発声や雑音データを含む音声を用いて識別関数を設計してきたが、限度があると考えられる。これからの方向性としては、個人性や雑音の性質によって個人差や雑音を分類し、それぞれに対してモデルをあらかじめ用意しておき、認識の際には音素モデルと融合することによって認識することが主流になるであろう。

したがって、個人差、調音結合、発声速度、雑音などのモデルを組み込める識別関数を利用した音素モデルの設計法が重要である。当然時間的な伸縮も考慮できる必要がある。さらに、これらの音素モデルが時間方向に並列に並んだ2次元的なモデルにおける、最適なネットワークアーキテクチャとその設計法が必要である。このアーキテクチャでは時間的に同時の音素のみならず、時間的に近いものや遠いものとも結合があり、認識の際は時間的に早い音素モデルから発火し、それが時間的に後続する音素モデルへとその影響が減衰しながらも伝搬していくことが必要と考えられる。このような性質を持った音素モデルとアーキテクチャの設計が主な問題である。

4.2.3 サブ問題

4.2.2の問題は以下のサブ問題を含む。

(1) 全体の最適性を考慮した個々の識別関数の最適設計法、あるいは逆に個々の識別結果の最適統合法

(2) 効率の良い音素コンテキスト依存型の識別関数設計法

(3) 時間的伸縮を許容する多数個の識別関数の設計法とそのためのクラスタリング手法

(4) 個人差、調音結合、発声速度などに柔軟に対処できる識別関数設計法

(5) 複数個の識別関数を階層的、時間的に統合し、かつその結果が伝幡して行くネットワークのアーキテクチャ設計法

4.2.4 発展問題

いままでの議論は音素や音節の知覚レベルを対象としたものであるが、発展問題として、次のものが考えられる。

(1) より高次の言語知識との情報の相互伝幡、音

素対の識別関数、個人差などの種々の変動に対処できる、TRACEモデルに類似したネットワークアーキテクチャの設計理論。

(2) 前述のネットワークアーキテクチャのハードウエア実現論。　　(牧野正三)

4.3 ディクテーションマシン

4.3.1 背景

(1) ディクテーションマシンの必要性

ワードプロセッサの普及により、文書の作成においては専門のタイピストに頼らず、自分でワープロを打つことが自然になってきた。文書のフォーマットが確定してくると、必要事項だけ記入すればよいようになり、文書の作成は大変効率がよくなってくる。ブックタイプのワープロともなれば、いつも持ち歩いていて、思いついたことを書き留めておき、時間のあるときにまとめるということも可能である。

しかし、もっと便利なものを求めるならば、持ち歩くことに抵抗がなくなった携帯電話や、携帯用カセットを使って、思いついたことを自由にしゃべり込んでおき、後で秘書に掘り起こしを頼んで原稿にする方法であろう。これは、口述筆記と呼ばれているものであり、欧米の会社では秘書を雇える立場にある人がよく使っている方法である。

もし、自由にしゃべり込んでおいた内容が、人手を介することなく機械で口述筆記できれば、時間の使い方や文書の作成効率は今よりもっと上がるものと思われる。何よりも、秘書が雇えなくても口述筆記が可能となり、新しいワープロとして世の中に広まるであろう。これがディクテーションマシン（口述筆記機械）である。現在でも音声ワープロは商品化されているが、これは、ワープロのキーボード入力を音声入力にかえたものである。

このディクテーションマシンと音声ワープロが結合して威力を発揮するのは、ペンポイントタイプの電子手帳が普及したときであろう。もはやキーボードはないので、指示はすべて音声かペンである。長い文書や思いついたアイデアは、音声でしゃべり込んでおくと後で口述筆記されて蓄積されるであろう。電子手帳が更に小さくなり、ディクテーションマシンが腕時計の中に埋め込まれ、通信機能を使ってデータを転送することができるようになれば、一人一台ディクテーションマシンを持つ時代が到来するであろう。

(2) 現在の到達点

自由に連続発声した音声に対するディクテーションマシンはまだ存在していない。これに比較的近い音声ワープロでは、当初、音節毎に区切って発声した文章しか受けつけなかった。その後、文節発声の文章を受け付けることができるようになり、現在では、連続的に自由に発声した文章を入力できることが望まれている。

現在の音声ワープロの基礎になっている代表的な音声認識技術は、音素／音節の学習と認識技術である隠れマルコフモデル（HMM）である。HMMは、音素／音節の確率モデルを大量のデータから自動学習するものであり、学習したモデルを結合して単語モデルを構成できることから、大語彙連続音声認識の基本技術として用いられている。

ディクテーションマシンは、連続音声から、音素、音節、単語の系列を正確に生成することを目的にしている。従って、音素／音節、単語のHMMと、最適な音素／音節、単語系列を探索するビタビ・アルゴリズムを連続音声に適用することにより、ディクテーションマシンの基本を構成することができる。

しかし、音素／音節、単語の正確な系列を連続音声から生成するとなると、基本技術だけでは不十分である。精度を上げるためには、音素／音節、単語間の制約を取り入れる必要がある。例えば、どの音素の次にはどの様な音素がくるかといった音素対や、それを確率で表現したバイグラム、更に3音素連鎖に拡張したトライグラムや、一般的にはNグラムを用いる必要がある。音素だけではなく、音節や単語においても、このタイプの統計量が必要である。特に単語におけるこの統計量は統計的文法と呼ばれている。これに対して、オートマトンや、LRパーザ、文脈自由文法といった言語学の分野で用いられている構造的な文法を用いることもできる。

(3) 現状と目標とのギャップ、問題点

オートマトンや、LRパーザ、文脈自由文法のような構造的文法を使用する場合には、生成規則の数を有限にしておくために、応用の場面（タスク）を限定する必要がある。これでは、自由な連続発声に対するディクテーションマシンを構成することができない。できる限り、統計的文法のような弱い制約を持った文法が必要となってくる。現時点では、音素／音節、単語のHMMの精度（HMMを用いた場合の、音素／音節、単語の認識率）は決して高くはなく、統計的文法の制約もそう強くはないため、精度の良いディクテーションマシンを構築することは難しい。

また、発声の単位が大きくなるにつれ、音声パターンの変形が強くなるという問題もあり、発話速度に依存しない認識方法、音素コンテキストを取り込んだパターンの変形に強い認識方法などを開発しなければならない。ディクテーションマシンが一般的な環境で使用されることを考えると、発話環境を推定し、そこでの雑音を認識して除去するなど、雑音に強い認識方法が必要である。

更に、「あのー」、「えー」といった文の意味に直接影響を与えない不要語を処理する方法や、辞書に登録されていない未知語を検出する方法の研究などが必要である。また、言い淀みや中断が生じた場合には、文の意味理解も必要になる。これには、一文一文の理解だけでなく、文と文との意味的な関係などの理解も必要であると考えられる。

4.3.2 主な問題

(1) 自由に発声された連続音声から、単語の辞書を用いずに、音素・音節間の制約を用いて、音素・音節系列を生成せよ。

単語の辞書とは、各単語がどのような音素系列で構成されているかを表わしたものである。この問題は、単語の知識を用いることなく、音素・音節間の制約に関する文法のみを用いて、連続音声から音

素・音節系列を生成する問題である。どの様な文法を音素・音節間の制約にまでくだいて利用するかが研究の中心となる。

(2) 自由に発声された連続音声から、100～1,000単語の単語辞書、単語間の制約、音素／音節間の制約を用いて、既知語（辞書に登録されている単語）の系列を生成せよ。また、既知語以外の単語については音素／音節系列として生成せよ。

自由に発声された連続音声において、100～1,000単語といった少ない単語間の制約を、バイグラムやトライグラムのような統計的文法や、文脈自由文法のような構造的文法から求めることは難しい。単語間の制約が使えない場合、キーワードを単独で検出するキーワードスポッティングを行なうのが一般的である。キーワードスポッティングでは、キーワードのみを探索するのではなく、連続音声に対して最適なキーワード、音素／音節系列を生成するといった観点が必要である。(1)の問題を単語を含むように拡張することによって実現できるが、辞書に登録されていない未知語や不要語を判定して、それらを音素系列として出力する方法が研究の中心となる。

(3)自由に発声された連続音声から、1,000～10,000単語の単語辞書、単語間の制約、音素／音節間の制約を用いて、既知語の系列を生成せよ。また、既知語以外の単語については音素／音節系列として生成せよ。

単語辞書の語彙数が多いので、単語間の制約をバイグラムやトライグラムのような統計的文法や、文脈自由文法のような構造的文法から求めることができる。文法を用いて連続音声から既知語の系列を生成する場合、文法にのらないしゃべり方、すなわち、文法の不完全性をうまく扱う方法が必要になる。また、(2)と同様、未知語や不要語に対する言語的扱いや、音声としての扱いなども必要となる。連続発声の文章から部分的に認識された単語と、辞書中の単語間の関係から発話の内容を理解し、これを基に認識されていない未知語や不要語の意味を推定することが可能となる。例えば人名や地名など、辞書にない単語は音節系列として出力することなどができるようになる。

4.3.3 応用例

会議において、議事録を自動作成する。これには、マシーンを話者に適応させることや、雑音を除去する事などが含まれている。更に、議論が白熱すれば複数話者が同時にしゃべった内容の聞き取りなどが必要になる。一部聞きとれないところは、意味理解をして補って書きとることも必要である。

（有木康雄）

4.4 実時間音声会話アミューズメント・システムの構築問題

4.4.1 背景

音声は、実世界の中でそれが話される場合、その内容についてかなり冗長な部分を含んでいる。音声を用いて計算機と会話する場合も、人間側の音声の発声については同じように冗長とするときが自然で人間に負担が少ない。すなわち、会話音声を計算機が理解の対象とするとき、人間の話す音声を逐一認識する必要はないと考えられる。しかし、このときでも人間と計算機の間で音声の会話が継続するかぎりにおいて、会話の意味の主要な内容は音声を通じて理解されていることが必要である。音声理解の機能について、最低限の会話継続性の可能性からそれを評価することは、音声理解の研究において新しい展開をもたらすことになると予想される。ここでいう最低限の会話継続の定義とは一般に明確ではないが、ここではそれを計算機との音声による会話が、人間にとっての娯楽の一種として位置付けられうることとして定義しよう。

4.4.2 主たる問題の内容

音声による計算機と人間との会話のモデルを設定するにおいて、人間側が単数である場合と複数である場合とするものがある。より興味あるものは人間側が複数である場合であろう。いま、複数の人間が会話をしている状況で、その会話の話題を計算機が理解するということを想定する。そのとき、主たる問題とは、例えば２人の人間の会話に計算機が参加することによって、３者による「会話」が継続し、それが不自然でなく継続し、人間側にとってそれが一種の娯楽として会話自体が捉えられるという状況を実現することである。

本構築問題を解くことができるシステムの特徴としては以下のものが挙げられる。

(1) 人間側の自由な会話音声をともかく許容する機能をもっている。
(2) 娯楽性のある実証実験となっている。
(3) 話題の変化する談話モデルを構築している。
(4) 音声理解システムを実時間で動作している。

(1)は、人間側（複数）の会話の進行に、計算機によるその会話の理解の結果が不自然さもたずしてとりこまれることを要請する。計算機による理解状況を表示するものは、言語であったり画像であったり、合成音声による出力であることが想定される。人間側の会話は、これらの計算機による理解状況の表示および応答と、密接さを保って続行される。

(2)は、単にある種のタスクを音声コマンドによって実行させるのではなく、音声を用いることによって生じる娯楽性のある会話を可能とすることの要請である。

(3)は、話題の持続性、変化について会話モデルが破綻せずに適応できることを要請している。この機能は、音声理解の中で、扱える単語の数、文構造の種類の複雑さに関連した部分と、会話モデル自体の学習機能に関連した部分、とに分けられる。

(4)は、音声会話理解システムの、人間を含む会話への参加の仕方に関し、人間の間での会話の途中への割り込みなどが自然に行なわれることを要請している。

4.4.3 サブ問題

以下で述べるサブ問題は主たる問題の部分問題の例である。

(1) 会話における娯楽性を保証する心理的研究を行なうこと。

(2) 音声理解の対象になる文集合の動的決定機構を明らかにすること。

(3) 何を外部に問い合わせるかを同定する推論メカニズムの解明。

(4) 計算機が把握している会話状況の説明機能をもたせること。

(1)は、音声を用いる会話がどのような特徴をもてば、娯楽性を獲得できるかを心理的にも明らかにする。

(2)は、会話の継続のために、各時刻における文脈依存による文認識候補の範囲をモデル中に絞り込む方式を明らかにする。

(3)は、会話理解システムにとっての統一性の確保と、前記文認識候補の数の減少のために、どのような部分を確定することが求められるかに基づいて、外部に問い合わせる部分を質問文という形で作成する。

(4)は、会話へ参加する人間にとって計算機がどの程度の理解状況にあるかを知らせるために、理解状況を表現する別個の手段を設けることを意味する。

4.4.4 発展問題

主たる問題の解決の試みをする中で、より発展した問題が浮かび上がってくると思われる。それらを列挙すると以下のものとなる。

(1) 音声会話理解と画像理解を組み合わせることによる理解機能の向上。

(2) 大規模データベースの検索と音声会話理解機能との統合。

(3) 音声会話理解システムによる会話におけるリード役を積極的にとれるように機能の発展を図る。

(1)は、会話理解状況の説明・表示に言語的記述のみではなく、画像を用いることや、計算機による画像理解の結果と音声会話理解とを結びつけることによる、相互補間を行なうことを意味する。

(2)は、理解システムが単にモデルをもつということにとどまらず、大規模データベースにアクセスできれば、より高度で広範囲の知識を議論の対象としうることを意味する。

(3)は、音声会話において話題を提供、リードする主体を計算機による理解システムとし、人間側の思考の発展を刺激する存在となることを意味する。

（岡　隆一）

4.5 音声認識と自然言語処理との融合

4.5.1 背景

音声認識での言語処理の役割は、言語情報によって次に認識すべき単語や音韻をできるだけ限定するいう観点から研究が行なわれてきた。このような流れに合致するモデルとして、単語や音節や文字のトライグラムが用いられてきた。現在の音声認識での言語モデルの研究の流れは、大量のテキストデータベースに基づく統計的な言語モデルの構成に重点がおかれている。しかし一方で、このような言語モデルの骨組みとして、有限オートマトンや単なるトライグラムでなくて、文脈自由文法を用いようとする研究もふえて来ている。特に、一般化予測ＬＲパーザが広く使われ始めている。さらに、チャートパーザやユニフィケーションなどを用いようとする小規模な試みもある。さらに、一歩踏み込んで、大量のテキストデータベースから文脈自由文法を学習しようとする試みや、言語モデルをユーザや特定のタスクに適応させる研究も開始されている。

言語モデルの予測能力の評価には、パープレキシティという情報量的な尺度が用いられてきた。

4.5.2 主な問題

音声認識の言語モデルが、単に単語等を予測するために用いられてきたが、自由発声等を取り扱うたには、発声内容を理解することが重要で、一音一音正しく認識することは、必ずしも必要ではない。また、発声内容の理解のためには、自然言語処理で行なわれているユニフィケーション等の手法を導入する必要がある。

音声は基本的に対話の手段として、人が最も使い易いメディアである。特に、機械に対しても音声で入力できることが重要である。よって、機械との対話という観点で音声認識の問題をとらえる必要がある。そのためには、機械の得意とするメディアを明確に意識して、対話システムを構成するという観点からの研究が必要がある。出力に関しては、機械は人と異なり、音声出力でなくむしろ画像や図形の出力が得意と考えられる。このようにメディアを意識して、音声認識の研究を進めることが、重要となってくる。

以上の観点から、音声認識の言語処理の問題をまとめる。

(1) 意味理解を考えた音声認識のための統計的言語モデル。

(a) 意味を基本としたパープレキシティの提案

自由発声音声は、不要語や冗長な表現を含んでいるために、従来のパープレキシティからすると、キーワードを連続単語発声したものよりパープレキシティが小さくなることがある。しかしながら、音声認識は連続単語発声より格段に難しくなる。このような矛盾に対して、意味を基本としたパープレキシティを考える必要がある。

(b) 意味を考慮した統計的言語モデル

同じ意味をもつ文でも、様々に表現される。意味と表現の多様性を考慮して、テキストデータベースから統計的言語モデルを学習することが重要となる。

(c) 文脈自由文法にユニフィケーションを組み込んだ効率のよい単語予測言語モデル

ユニフィケーションは、キーワードの間の制約を記述する優れた手法である。この手法とＬＲパーザなどの優れた予測モデルとを融合した、音声認識

に適した効率のよい単語予測モデルが必要となる。

(d) 対話状態に応じた言語モデルの適応化
音声対話システムでは、ユーザと機械の対話状態に応じて言語モデルを適応的に変更する必要がある。この対話状態に応じて言語モデルを適応化させるアルゴリズムを考える必要がある。

(e) 自由発声テキストデータベースの収集とannotation
上記のような音声認識、対話システムのための言語モデルを研究するためには、大規模な自由発声の音声および書き下したテキストデータベースが不可欠である。さらに、このテキストデータに対しては、様々なレベルでのラベル（annotation）が必要となる。

(2) 音声対話モデル

(a) 機械の特質を考慮した音声対話システムの構成
マルチメディアのパラダイムで、人の特質と機械の特質を理解して、メディアとしての音声が真に使いやすい、音声対話システムの構築を試みることが望まれる。

(b) 機械との対話を実現するための対話テキストデータベースの収集とannotation
人と人との対話と、機械と人の対話とは大幅に異なる形態になると思われる。マルチメディアのパラダイムでの機械との対話の、対話テキストデータベースの収集とannotationが重要かつ不可欠となる。

(c) 音声対話モデルの評価尺度
音声対話システムを評価するには、音声認識率や認識速度の他に、音声対話モデルを評価する尺度を研究する必要がある。

4.5.3 サブ問題
上記の問題は、雑音処理やユーザへの適応の問題の音声認識固有の問題のほかに、次のようなサブ問題を含む。

(1) マルチメディアの中で音声の果たすべき役割を明確にする。
マルチメディア、マルチモーダルのパラダイムのなかで、音声メディアの役割を明確にする。

(2) 音声入力時のマイクやAD変換のON／OFFをなくする手段の検討。
音声入力を真に使いやすくするには、音声入力の際のON／OFFをなくし、かつ、マイクの装着を不要にする必要がある。もちろん、マイクから離れたところで発声しても認識できるような、音響処理技術も検討する必要がある。　（鹿野清宏）

4.6 話者認識

4.6.1 背景
音声波に含まれる第一義的な情報は、言葉の意味内容に関連した音韻性情報であるが、第二義的情報として重要なものに、誰が話しているかという個人性情報がある。音声波から個人性情報を自動的に抽出して、誰の声であるかを判定することを話者認識という。

話者認識の研究は、犯罪捜査とも関連があるが、今後、情報化社会の発達に伴って、音声を個人のIDとして用いた、新しいサービスの需要が急激に伸びてくると予想されている。音声を個人のIDとして用いることができれば、鍵やカードを用いるよりも簡単であり、両手がふさがっていても使いことができ、盗まれる心配がないなど、メリットが多い。特に電話を用いたサービスでは、音声以外の情報は使えないので、音声で誰かということが判定できれば、特定のメンバーに限定された情報を提供したり、その本人に料金を請求することができるようになり、多くの新しいサービスの実現の鍵になると予想される。いたずら電話を撃退し、個人のプライバシーを守る手段にもなると考えられる。

話者認識に関する最も重要な課題は、同じ人の声でも、発声のたびに変化し、回線、マイクロホンなどの特性、雑音、などによっても変化するため、それが許容できるようにすると同時に、他人の似た声で誤動作しないようにすることである。

話者認識には、話者識別と話者照合がある。前者は、あらかじめ登録されている誰の声であるかを判定するものであり、後者は、音声と同時に自分が誰であるかのIDを入力して、その音声が本当にその人の声であるか否かを判定するものである。話者識別の誤り率は、方法が同じでも、登録話者の数が増えるにつれて単調に増加するが、話者照合の誤り率は、登録話者の数には依存しない。

話者認識の方法はさらに、発声すべき言葉（キーワード）をあらかじめ決めておく発声内容限定（依存）形と、任意の言葉を発声すればよい発声内容独立形に分類できる。発声内容限定形の場合は、標準パターン（あるいはモデル）と入力音声の時間軸を合わせて比較すればよいので、発声内容独立形よりも比較的容易である。発声内容独立形の場合は、発声内容による変動と、話者による変動を分離する必要がある。この問題と、音声認識における個人差への対処の問題は、音声波における個人性情報と音韻性情報の分離という意味で、表裏一体をなしており、共通したアプローチからの研究を進める必要がある。

話者認識における重要な問題の一つは、同じ人が同じ言葉を発声しても、そのつどその音声の物理的特徴が変化し、発声内容の認識の場合よりも、その影響が大きく現われることである。このため、話者認識のための特徴パラメータや種々の方法の評価実験を行なう場合には、必ず複数の異なる時期に収集された音声を用いて、学習用の音声と評価用の音声が異なる時期に収集されたものであるようにする必要がある。同じ時期の音声の変動は、異なる時期の音声の変動に比べて一般に極めて小さく、その違

いは特徴パラメータによっても違うので、同じ時期の音声だけを用いたのでは、正しい評価をすることができない。

4.6.2 主たる問題

話者認識に関して、今後解決すべき課題としては、次のようなものがある。

(1) 長期間にわたって安定で、しかも作り声や風邪などに強い特徴パラメータの探求

(2) 発声内容に依存せず、何を発声しても共通に含まれる個人性情報の抽出法の開発

(3) 任意の言葉を発声したときに、各音韻や音節に依存した個人性情報を自動的に抽出する方法の研究

(4) 音声認識の場合と同様に、電話機や伝送路による歪みや雑音、周囲雑音等を除去する方法の開発

(5) 種々の要因による特徴パラメータの話者内変動を自動的に正規化して、話者照合における判定（本人として受理するか、他人として棄却するか）のしきい値を安定に設定する方法の開発。

4.6.3 サブ問題

(1) 音韻や音節によって、個人性情報がどのように異なって含まれているか、どのような音韻や音節が話者認識により有効であるかを明らかにすること。

(2) 発声方法（発声速度、声の大きさ）などを変えたときにも安定している個人性情報を明らかにし、それを抽出する方法を開発すること。

(3) 人間の聴覚による話者認識では、どのような音声の特徴が主に用いられているのかを解明すること。

4.6.4 発展問題

音声認識と話者認識が同時に可能であるような、個人性と音韻性を同時に精度よく表現した音韻モデル（HMMなど）を、その話者の限定された量の学習音声（複数の文章音声など）を用いて構築する方法を考案すること。不特定話者の音韻モデルなどを、その過程で組み合わせて用いてもよい。

（古井貞熙）

4.7 音声言語の識別

4.7.1 背景

音声情報には、音韻（言語）情報以外に、話者の個人性情報、感情などの韻律情報、などがある。音韻情報を正確に抽出・認識するためには、その発話の言語（日本語とか英語）が何であるか同定しなければならない。パターン認識システムは認識対象毎にアーキテクチャが異なるのが普通である。例えば、音韻体系、音韻配列構造、文法などは言語毎に異なる。通常は、これは既知として扱われているが、そう仮定できない場合もある。この延長として、音声言語と非言語音の弁別の問題もある。音声言語識別は、多言語間通訳システムの入力言語が何であるか判定する場合とか、見知らぬ人が聞き慣れぬ言葉を発声した時の対応などの場合に有用である。また、音声による言語識別の研究は、話者情報を分離し音韻情報を抽出するという音声認識そのものの他にも、言語間や方言間の発生系統木や相互の特徴などの解明に役立ち、音声・言語学、言語教育学に寄与できる。

4.7.2 主たる問題

音声言語の識別問題とは、不特定の人が、自分の第一獲得言語の任意の文を発声したとき、それを聞いてその言語を識別することである。最も精度がよいと思われる音声言語の識別法は、言語ごとに不特定話者の連続音声認識システムを構築することである。しかし、これを実現するのは困難であるし、欠点もある。例えば、不特定話者の連続音声認識システムを構築するためには、大量の音声データベース、言語データベースが必要であること、また、たとえこのようなシステムが構築できても、言語識別システムが大きくなり、コストが高くなるなどの問題がある。

そこで言語毎に異なる事前知識（音韻体系など）を用いないで、最もよいと思われる音声言語の識別法を研究する。理想的には、音声情報を音韻情報、話者情報に分離する手法を解明し（交互作用もあるから完全な分離はできないが）、その音韻情報を基に、言語の音韻空間を構成し、その空間上での遷移（音韻配列）構造を究明すればよい。まず、基本問題として、以下の順で考察し、言語識別の研究をする。

(a) あらかじめ決められたテキストを発声してもらい、その発声者がその言語を母国語とするかどうかを判定するシステムはテキスト依存型の言語識別法である。この方法を考察する。

(b) 文字言語がアルファベット系列で表現されているとして、これを用いて言語識別する方法を考察する。ただし、音声では、単語間の区切り記号の発声はないから、空白記号は用いず、また単語辞書の情報を用いない方法を考察する。

(c) 音声言語で言語識別する場合、文字言語で言語識別する場合と共通する点と異なる点を考察する。

(d) テキスト独立型の話者認識と音声による言語識別の技術上の共通点と相違点を考察する。

(e) 音声をエルゴディック情報源とみなすとき、音声情報源の代表的なモデル化法は下表のように分類できる。この表をもとに、よりよいモデル化法とそれぞれによる言語識別法を考察する。

表4 音声情報源の代表的なモデル化法

離散情報源	記憶なし	コードブック （VQ歪み、出現頻度分布）
	記憶あり	VQ歪みHMM 離散出力分布型HMM 確率文脈自由文法
連続情報源	記憶なし	混合連続出力分布
	記憶あり	連続出力分布型HMM 連続出力分布型確率文脈自由文法

4.7.3 サブ問題

より精度よいシステムを構築するためには、話者によらない正確な音韻空間の構成と音韻識別（クラスタリング）および音韻配列構造をとらえるモデルを構築する必要がある。すなわち、次のサブ問題を研究しなければならない。

(a) 音声時系列パターンの時間軸上でのクラスタリング、例えば有声音と無声音の弁別、セグメンテーションなどの研究をする。

(b) HMM（隠れマルコフモデル）や従来のニューラルネットワークの欠点を考察し、それを補うモデルを研究する。特に動的な特徴、状態遷移情報などをモデル化する方法を研究する。

(c) 話者によるパターン変動に対処するためには

・多数の話者の音声データを統計的に処理する

・音声認識で用いられている話者正規化・適応化法を用いる

・話者に独立な音響的・言語的特徴を用いる

などが考えられる。それぞれの併用法も含めて言語識別法を研究する。

4.7.4 発展・応用問題

(a) サブ問題(a)(b)を発展させ、超極低ビットレート（100～150 bit/s）の音声符号化法および復号化法を研究する。

(b) 言語識別法を多言語の発音学習に利用する方法を研究する。

(c) ある言語の大量のラベリングのない音声データから、音韻（音節）体系、音韻（音節）配列構造などを自動獲得する方法を研究する。

(d) 音声と、雑音、非音声（動物の鳴声）とを弁別する方法を研究する。 （中川聖一）

5 あとがき

究極の音声認識・理解の実現を目指して、今後新たに展開すべき研究課題について述べた。音声は、人間にとって最も容易かつ便利な情報伝達・交換メディアであり、今後のマルチメディア環境による高度情報化社会においても、その重要性は変わらないであろう。音声認識・理解には、パターン認識・理解の一部として、さらに言語学、心理学、生理学、信号処理論、などとの境界領域に存在する分野として、学問的にも興味深い課題が無数に存在している。ここで紹介したのは、その中の代表的な部分である。今後の音声認識・理解研究の大きな飛躍のために、多くの研究者・技術者の方々が参画して下さることを期待したい。

参考文献

[1] 古井貞熙：ディジタル音声処理、東海大学出版会(1985)

[2] 古井貞熙：音響・音声工学、近代科学社(1992)

[3] L. R. Rabiner, B. H. Juang: *Fundamentals of Speech Recognition, Prentice Hall* (1993)

[4] S. Furui: "Towards the ultimate synthesis/recognition system", *NAS Colloquium on Human-Machine Communication by Voice* (1993)

[5] S. Furui: "Toward robust speech recognition under adverse conditions", *ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandeliew*, pp.31-42 (1992)

[6] 中川聖一：確率モデルによる音声認識、電子情報通信学会(1988)

[7] 甘利俊一監修、中川聖一、鹿野清宏、東倉洋一：音声・聴覚と神経回路網モデル、オーム社(1990)