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)	音声認識の過去・現在・未来
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	NTT R&D, Vol. 43, No. 10, pp. 1055-1060
Citation(English)	, Vol. 43, No. 10, pp. 1055-1060
発行日 / Pub. date	1994,

Powered by T2R2 (Science Tokyo Research Repository)

音声認識の過去・現在・未来

Speech Recognition —— Past, Present and Future ——

古井貞熙*
Sadaoki FURUI

あらまし

本論文では，音声認識，すなわち音声の意味内容の認識（狭義の音声認識）と発声者の認識（話者認識）に関する研究の過去・現在・未来について述べる．音声認識の研究の40年以上の歴史の中で，日本の研究者の果たしてきた役割は，米国に次いで大きい．**隠れマルコフモデル**†（HMM），**Δケプストラム**†，**統計的言語モデル**† などの利用によって，不特定話者が発声した大語彙の連続音声の認識性能は，近年大幅に向上している．話者認識技術も，HMMの利用，**尤度正規化法**†，テキスト指定型の提案などによって精度と信頼性が向上している．NTTのANSERシステム，AT&Tのオペレータサービスの自動化システム，Sprint社の音声テレホンカードなどの実用システムでは，すでに多数の利用者を得ている．しかし，現在の音声認識技術でできることは，人間の能力に比べると極めて限られている．音声の境界領域的性格に立脚した新たな研究の展開が求められている．

Abstract

This paper presents a brief history, then describes the current state of the art and future prospects for speech and speaker recognition technology. The recent incorporation of several new techniques —— such as the hidden Markov models (HMMs), Δcepstral parameters, and stochastic language modeling —— has greatly improved the accuracy of speaker-independent, large-vocabulary, continuous speech recognition. Speaker recognition accuracy has also been improved by using the HMMs, likelihood normalization, and the text-prompted method. However, we still need continued research on the inter-disciplinary aspects of speech before systems approaching human capability can be achieved.

まえがき

音声認識とは，人が発声した音声を知識処理することである．音声認識は，発声者が意図した言葉の意味内容の認識（狭義の音声認識）と，発声者が誰であるかの認識（話者認識）に分けることができる．音声認識システムを実現するための研究は，1950年代から今日まで，世界中の多数の研究者によって，40年以上にわたって行われてきた．著者自身がこの研究分野に足を

*NTTヒューマンインタフェース研究所　NTT Human Interface Laboratories
†論文中のゴシック文字は，本論文末に解説があります．
©日本電信電話株式会社　1994

踏み入れてからも，もうすぐ四半世紀が経過しようとしている．その間，Stanley Kubrickの有名な映画「2001年宇宙の旅」の中のコンピュータHALや，George Lucasの「スターウォーズ」のシリーズの映画の中のロボットR2D2を始めとして，音声認識のできる数多くのロボットや装置が，空想科学の中で人々の関心を集めてきた．話し言葉を認識し，その意味を理解することができる知的機械を作りたいという願望は，多くの研究者の夢であり続けてきた．しかしながら，これまでの40年間の大きな技術的進歩にもかかわらず，任意の話題についての，あらゆる人による，あらゆる環境での音声を理解できるという目標からは，まだ程遠い状況にある．

音声は目で見ることができないので，音声認識の難しさを直感的に把握するのは難しいが，いわば毛筆の続け字で書いた行書あるいは草書をコンピュータで読み取ったり，誰の筆跡であるかを判定するようなものである．個人差や雑音（よごれ）によるパタンの変形が大きく，書き間違いに相当する言い間違いや言い直しも，日常の会話では頻繁に起こる．これらのすべての問題に対応することはとうてい不可能であるが，ある程度は対応できないと使い物にならない．狭義の音声認識と話者認識を比べると，研究者の数は圧倒的に前者の方が多く，両者には難しさが異なる面もあるが，共通した技術や課題も多い．本論文では，両者に関する研究の歴史と現状を簡単に紹介し，将来の展望を述べる．

音声認識研究を振り返って

2.1 音声の意味内容の認識（狭義の音声認識）

音声認識システムに関する世界で最初の論文は，1952年に米国のベル研究所のDavisらによって発表された「Audrey」システムに関するものである[1]．このシステムは，一人の話者の，区切って発声された一桁数字音声を認識するものであった．その後，同じ1950年代に，米国のRCA研究所で単音節を，英国のUniversity Collegeで音素を認識するシステムの研究が行われた．

後者の研究の重要な点は，単語中の音素の認識精度を上げるために，英語における可能な音素連鎖に関する統計的情報を用いたことで，これは現在の主流になっている統計的言語モデルのさきがけであった．不特定話者の音声を認識する最初の試みは，1959年に米国のMITのLincoln研究所で作られた母音認識装置であった．

1960年代になると，電波研究所（現在の通信総合研究所），京都大学，NEC，日本電信電話公社（現NTT）研究所（以下，通研と略記）などの日本の研究機関が音声認識に参入し，母音などの音素や数字音声の認識に関して優れた成果を上げた．1960年代の終わりから1970年代の初めにかけての重要な研究成果として，動的計画法（DP：dynamic programming）による時間軸の整合と，線形予測分析法（LPC：linear predictive coding）がある[2]．前者は，音声の時間軸の伸縮に対処しながら類似性を計算するための極めて効率的な方法であり，NECとソビエト連邦（現ウクライナ）でほぼ同時に独立に提案された．その基本的考え方は，現在の音声認識でも広く用いられている．LPCは，音声生成のモデルに基づいた優れた分析法として，通研の板倉と斎藤（敬称略）によって提案された．板倉は，1970年代の初めに米国のベル研究所に客員研究員として滞在していた間に，この二つの成果を組み合わせて高い認識性能を示し，ベル研究所においてしばらく中断していた音声認識研究を再開させる重要なきっかけを作った．

1970年代のその他の重要なマイルストンとして，NECで提案された二段DP法を初めとする連続単語音声認識のためのアルゴリズム，米国のCarnegie Mellon大学（CMU）やIBMにおける連続音声認識の研究，ベル研究所における不特定話者の音声を認識するための技術，通研で提案された音素や音節を単位とする認識アルゴリズムなどがある．

1980年代は，それまでのテンプレート（標準パタン）との照合を基本とする方法から，隠れマルコフモデル（HMM：hidden Markov model）を代表とする統計的モデル化に基づく方法への移行によって特徴づけることができる．

HMMによる方法は，IBMなど少数の研究所ではそれ以前からよく知られていたが，一般的には，1980年代の半ばにその具体的方法と理論が広く発表されるようになって，初めて世界の研究機関で用いられるようになった．言語モデルに関しても，それまでの生成規則による方法に代って，統計的方法が用いられるようになり，IBMにおいて単語の連鎖確率による方法が盛んに研究された．

1980年代はさらに，米国のDefence Advanced Research Projects Agency (DARPA, 現ARPA) による1,000単語規模の連続音声認識プロジェクトが，米国における音声認識研究の原動力として，大きな役割を担うようになった年代であった．その傘下で，SRI，MIT，CMU，BBN，MIT Lincoln研究所などにおいて，熾烈な競争を背景に重要な研究成果が数多く生まれた．その一つは，大規模な音声言語データベースの構築と，それに基づく精密な音素モデルや統計的言語モデルの利用である．一方，日本では関西に，通研からの出向研究者を中心とするATR (㈱国際電気通信基礎技術研究所) の音声認識グループが発足し，自動翻訳電話を目標とした研究が活発に行われるようになった．1980年代の，世界に先行した日本の特徴的な研究として，ATR，信州大学，通研などにおける**話者適応化技術**†が挙げられる．

2.2 音声の発声者の認識 (話者認識)

話者認識には種々の応用が考えられるが，最初は犯罪捜査への利用であった．1660年の英国チャールズI世の死に関する裁判のときに，音声が犯人の割り出しの手掛かりとして初めて用いられたと言われている．しかし当時は人間の耳に頼っており，音声の個人性に科学的なメスが入れられたのは，1930年代のLindbergh氏子息誘拐事件のときが初めてである．1940年代には，ベル研究所のPotterによってサウンドスペクトログラフが発明され，いわゆる声紋 (voice print，visible speech) が自動的に描かれるようになった．1960年代になって，ベル研究所のKerstaが，これによる話者認識の可能性を示し，米国の裁判で初めて証拠として用いられた．

自動的な話者認識に関する論文は，ベル研究所のPruzanskyらによるものが初めてで，1960年代のことである．その後，1970年代初めにかけて，IBM，TI，通研などでも研究が開始され，テキスト独立型と依存型の両方に関して研究が進められた[2]．通研では当初から，ケプストラムパラメータの優位性に着目し，世界に先駆けてこれを用いてきた．1970年代の終わりには，ベル研究所に滞在していた著者らが，この時系列の**多項式展開係数**†を動的特徴として組み合わせる方法を試み，高い認識性能を示した．このケプストラムとその多項式展開係数（現在では，Δケプストラム，ΔΔケプストラムなどと呼ばれている）を組み合わせて用いる方法は，1980年代になって，通研で（狭義の）音声認識に初めて適用された．その優れた性能のため，この方法は現在世界中のほとんどの音声認識システムで用いられている．

話者認識の研究も音声認識と同様に1980年代から，HMMに代表されるような統計的方法を基礎とするようになり，精度の大幅向上が達成された．1980年代の重要な成果の一つに，米国ITTによる尤度正規化法の提案がある．これは，話者照合における判定のしきい値を安定に設定するために，発声時期や発声内容の変化による尤度値の変動を，尤度比の概念によって正規化する方法である．

3 現在の音声認識技術

現在実用化されている音声認識技術は，10年以上前に通研で開発されたANSERシステムのように，区切って発声された単語を認識するか，最近ベル研究所で開発されたオペレータサービスの自動化システムのように，文章音声の中からキーワードを検出（スポッティング）して認識する方法によるものである．現在では世界中の多くの機関で，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基本的な研究から，具体的な実用システムの実現を目的とした開発まで，広範囲にわたって活発に研究開発が行われている．

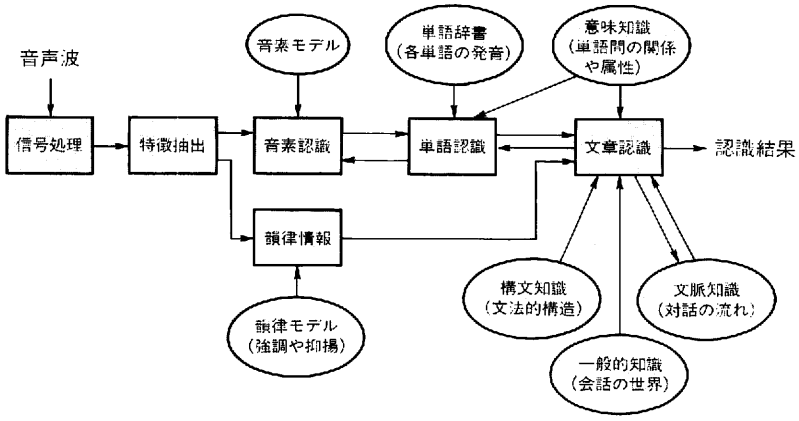

図1　現在の音声認識システムの典型的構成

現在研究されている大語彙の連続音声認識システムの典型的な構成を，**図1**に示す[3]．音声波は，信号処理部でディジタル変換された後，特徴抽出部でケプストラム，Δケプストラムなどのパラメータ時系列に変換される．声の高さ（ピッチ）や大きさの時間情報である韻律情報を取り出して，用いる試みも行われているが，まだ十分成功していない．一方，システムの側から，現在の話題の状況，言葉の意味，文法などに応じて，ユーザが発声する文（仮説）を予測し，その文を単語の系列で表現する．それをあらかじめ学習処理によって作られている種々の音素モデルの系列に変換する．各音素モデルは，HMMで表現することが多い．入力音声のパラメータ時系列が，この音素モデルの系列から出現する尤度（確率）を計算し，仮説としての文の言語的妥当性の尤度と組み合わせて，その文が発声された総合的な尤度（可能性）とする．次に異なる文（仮説）を作り，この（総合的）尤度を計算する．これを繰り返して，最も尤度の高い文を求め，その文が発声されたものと決定する．

このように現在では，言語的知識の主導による，いわゆるトップダウン的方法によって認識処理を進めることが多い．通研では，この原理に基づいて，80,000語の超大語彙を対象としたアルゴリズムの研究を進めている[4]．

現在実用化されている話者認識技術は，米国Sprint社のテレフォンカード（Voice Phone Card）のように，発声内容依存型のみである．しかしこれだけでは用途が限定されるため，現在，種々の目的を想定して，テキスト独立型や指定型の研究も行われている．テキスト指定型の方法は，録音した音声によって装置がだまされるのを防ぐために，認識のたびに新しい発声内容を装置の側から指定する方法で，最近通研で提案した方法[5]である．

音声認識技術の今後の展開

表に，音声認識システムの今後の実用レベルの発展の予想を示す．これはベル研究所のRabinerとJuangによるオリジナル[1]に，著者が手を加えたもの[3]である．今後の音声認識の主な用途としては，データベース（情報）検索・案内・操作，自動予約・注文，ディクテーション，音声ワープロ，電子秘書，ロボット，自動翻訳電話，身体障害者の補助機器などがある．

話者認識は，今後，電話を用いたバンキングや買い物サービス，情報提供サービス，コンピュータのリモートアクセス，クレジットコールなどにおいて，音声を本人確認の手段として用いる技術として，広く用いられるようになると期待されている．しかし，音声認識および話者認識の現在の技術でできることは，人間の能力に比べると極めて限られている．技術が実際に使われるためには，その技術をステップを踏んで進歩させるとともに，その現状レベル（能力と限界）にマッチした具体的応用を開拓するこ

表　音声認識システムの今後の発展の予測

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)	4～10DSP (200Mips/Chip)	5～50DSP (200Mips/Chip)	20～60DSP (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 自動翻訳電話

DSP：Digital Signal Processor

とが必須である．今後の技術的課題については，本特集の他の論文に詳しく述べられているが，最も重要な課題は，話し言葉の言語モデル（規則）を作り上げることと，人による声の違い，体調や精神状態，電話機や回線，周囲の騒音の付加などによる音声の変動の影響を受けない，頑健な方法を確立することである．

そのためには，単に音声現象の表面だけを見た方法ではなく，現象の本質をとらえた着実な研究開発を進める必要がある．音声認識の技術は，**図2**に示すように極めて多くの科学と工学の分野と関係しており，境界領域的色彩が強い，自然科学と工学の境界にある技術であるということもできる．音声認識技術を進展させるためには，これらの広い範囲の分野からの知識と技術を必要とする．もちろん，すべての分野の専門家になることは誰にもできないが，多くの分野の基礎をよく理解していることが必須である．

例えば，通常の文音声のように複数の音素や音節が連続して発声されるとき，舌，顎，唇などは，非同期，並列で，しかも相互に関連を持ちながら動く．現在の音声分析技術では，現象をスペクトルの時系列として一面的にしか見ないが，今後は音声生成機構に基づいて，隠れた複数の要因に分解して分析することが必要になろう．人間における話し言葉の理解過程の解明も，書き言葉とは全く異なった話し言葉の言語モデルの構築のヒントを得るためには不可避であろう．

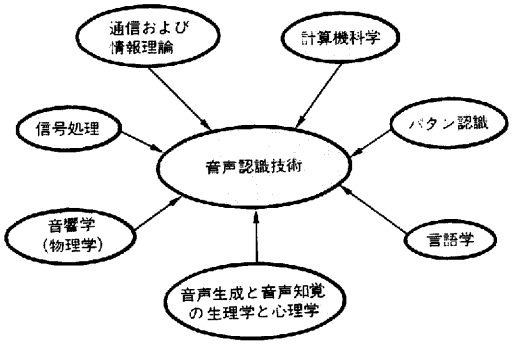

図2　音声認識に関連する科学と工学

あとがき

音声認識の技術は，40年の歴史を経て着実に進歩し，幾つかの具体的な応用で成功しているが人間並み，できればそれを超える能力を工学的に実現したいという我々研究者の目標からは，まだ程遠い状況にある．今後，進歩の正しい道筋を見失わないようにしながら，一層の努力を積み重ねていく必要がある．

本特集では，音声認識の音響処理技術，言語処理技術，および話者認識技術について，最新技術の内容を解説するとともに，外部の方に現状の分析と今後の展望を書いていただいた．その1つは，かつての同僚で通研における音声認識研究のリーダの一人であった，奈良先端科学技術大学院大学の鹿野清宏教授に，もう1つは，ベル研究所において音声認識研究を長年にわたってリードしてきたLaurence R. Rabiner部長とその部下であるChin-Hui Lee博士に執筆をお願いした．後者は，英語で書かれたものを著者と松岡が分担して日本語に翻訳した．本特集を通じて，音声認識の最新技術の概要を多くの方々にご理解いただき，今後の展開へご期待いただければ幸いである．

文　献

(1) L. R. Rabiner and B. H. Juang : Fundamentals of Speech Recognition. Prentice-Hall, New Jersey, 1993.
(2) 古井：音響・音声工学，近代科学社，1992.
(3) 小川編著：パターン認識・理解の新たな展開 ——挑戦すべき課題——，第3章，音声の認識・理解，54-91，1993
(4) 南・山田・鹿野・松岡：番号案内を対象とした大語い連続音声認識アルゴリズム，信学論，J77-A，No.2，pp.190-197，1994.
(5) T. Matsui and S. Furui : Concatenated phoneme models for text-variable speaker recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Minneapolis, II-391-394, 1993.

■ 用語解説 ■

隠れマルコフモデル

音素や単語の音声のスペクトル時系列を，確率状態遷移機械からの出力として表現する方法．通常は1つの音素を3状態程度の接続で表現し，状態間の遷移確率と，各状態あるいは遷移における種々のスペクトルの出現確率で，各音素や単語を特徴づける．音声スペクトルの変動を統計的に効率良く表現できる．

ケプストラム

対数スペクトルを逆フーリエ変換したもので，次のような特徴がある．①人間の聴覚の特性に近い対数スペクトルと線形変換の関係にあるので，人間の聴覚に合った判定ができる．②高次の係数はスペクトルの微細構造を，低次の係数はスペクトルの包絡を表すので，適当な次数で打ち切ることにより，比較的少ない数のパラメータで効率良く，滑らかなスペクトル包絡を表現することができる．

統計的言語モデル

連続音声認識において高い認識精度を得るためには，音声の音響的性質だけでなく，言語的制約すなわち言語モデルを利用することが必須である．近年，大量の言語データベースを用いて，単語や音素の連鎖関係をバイグラム（二連鎖），トライグラム（三連鎖）などの連鎖確率で表現することが広く行われ，認識精度の向上に大きく貢献している．

尤度正規化法

話者照合においては通常，未知の音声が，ある話者のモデルから生じたものである尤度（確率）を計算し，その値があらかじめ設定してあるしきい値よりも大きければ，その話者の音声であると判定する．この尤度の絶対値は，発声内容や発声時期による音声の変動のためにばらつくので，それ以外の話者のモデルによる尤度との相対値に変換する．これにより，値のばらつきが削減され，話者照合性能が大きく向上する．

話者適応化技術

音声認識において，話者によって声が違うことが認識性能を低下させる主たる原因の一つである．このため，話者が変わるごとに，その話者の限られた量の音声を用いて，各音素や単語のモデルをその話者に自動的に合わせる技術が研究されている．この適応化に用いる音声の発声内容が分かっている「教師つき」の方法と，任意の音声を用いることができる「教師なし」の方法が研究されている．

多項式展開係数

音声の各時点を中心とする50～100ms程度の区間のスペクトルパラメータの時系列を，多項式で展開したときの係数（微係数に相当）で，スペクトルの動的性質を表す方法．通常，スペクトル（包絡）をケプストラムを用いて表現し，その時系列の多項式展開係数を計算して，一次の係数をΔケプストラム，二次の係数をΔΔケプストラムと呼ぶ．

古井　貞熙

ヒューマンインタフェース研究所古井特別研究室長
昭和45年入社，主として音声認識，話者認識，音声知覚などの基礎研究に従事．53～54年米国ベル研究所客員研究員，平成元年ヒューマンインタフェース研究所音声情報研究部長，3年より現職．6年より東京工業大学客員教授．
昭和43年東京大学工学部計数工学科卒業．45年同大学院工学系研究科修士課程修了．53年工学博士（東京大学）．
IEEE Fellow・ESCA・電子情報通信学会・日本音響学会会員．
昭和50年電子情報通信学会米沢賞，63年および平成5年同論文賞，平成2年同著述賞，昭和60年および62年日本音響学会論文賞，平成元年科学技術庁長官賞，同年IEEE ASSP Society Senior Award各受賞．1993年IEEE Signal Processing Society Distinguished Lecturer.