論文 / 著書情報
Article / Book Information

論題(和文)	音声認識
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会誌, Vol. 78, No. 11, pp. 1114-1118
Citation(English)	The Journal of the Institute of Electronics, Information, and Communication Engineers, Vol. 78, No. 11, pp. 1114-1118
発行日 / Pub. date	1995, 11
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) 1995 Institute of Electronics, Information and Communication Engineers.

電子情報通信学会誌 Vol.78 No.11 pp.1114-1118 1995年11月

特集

2-7

2. 各分野における技術の変遷

音声認識

古井貞熙

古井貞熙：正員 NTTヒューマンインタフェース研究所

Sadaoki FURUI, Member (NTT Human Interface Laboratories, Musashino-shi, 180 Japan).

ABSTRACT

音声認識の研究には約40年の歴史があり，その中で種々の変遷を遂げてきた．線形予測分析法，Δケプストラム，DP（動的計画法），HMM，統計的言語モデルなどの技術は，統計的処理が中心となっている最近の音声認識の中で生き続けているが，過去に盛んに研究されたが統計的処理に合わなかったり，その枠組みの中では意味を失った技術も多い．音声認識技術は，まだ解決しなければならない課題も多いが，今後，マルチメディア環境の中で，種々の形で使われるようになると期待される．話者認識技術の変遷についても簡単に触れる．

キーワード：音声認識，話者認識，統計的処理，HMM

1. 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，発声者がだれであるかの認識（話者認識）に分けることができる．本稿では，主に前者を対象とするが，両者の技術の変遷には互いに関連があるので，適宜後者についても触れることにする．このため以下では，「音声認識」を狭義の意味で「話者認識」と区別して用いる．

2. 現在の音声および話者認識技術

現在の音声認識システムの典型的な構成を，図1に示す．大語彙連続音声認識では，ボトムアップに処理を行ったのでは，文章仮説（認識結果の候補）の数が爆発的に増えてしまう．このため，認識対象タスクに応じた言語モデル（単語辞書，意味情報，構文情報，文脈情報など）をトップダウンで用いて文章仮説を生成（予測）し，それを音素系列で表したものを逐次的に入力音声と照合して，検証していく方法をとる．

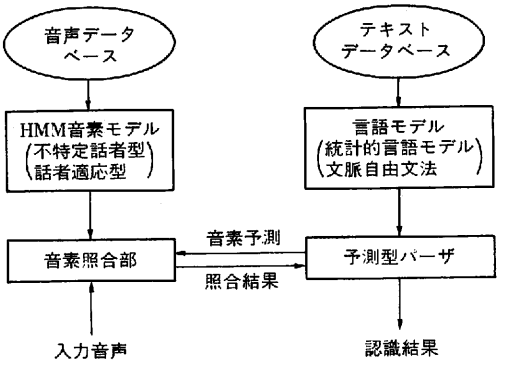


図1 現在の音声認識システムの典型的な構成　言語モデル（単語辞書，意味情報，構文情報，文脈情報など）をトップダウンで用いて文章仮説を生成し，それを逐次的に入力音声と照合して検証していく．

入力音声は，音素照合を行う前に5～10 msごとにスペクトルパラメータベクトルに変換（特徴分析）しておく．具体的には，世界中のほとんどのシステムで，ケプストラムとその時間変化の微係数であるΔケプストラムが用いられている．各音素の音響モデルとしては，隠れマルコフモデル（HMM：Hidden Markov Model）を用いる．各単語辞書は，単語の発音

を音素モデルの系列で表すものであるが，いわゆる調音結合現象のため，各音素の発音は前後の音素の影響を受けて変形する．このため，各音素ごとに前後の音素に依存した複数のモデル（音素環境依存モデル）を用いる．

言語モデルとしては種々の方法が研究されているが，最も広く用いられているのは，統計的言語モデルである．具体的には，バイグラム（bigram），トライグラム（trigram）などの単語の連鎖確率が用いられる．その他の言語モデルとしては，文脈自由文法，有限状態ネットワーク文法なども用いられている．すべての音響的および言語的制約を満足する最も可能性の高い文を探索するアルゴリズムの研究も極めて重要であり，フレーム同期ビーム探索，スタックデコーディング，A* ヒューリスティック探索などが用いられている．複数の知識源を統合する方法として，Nベスト探索法が，一つの理想的方法として広く用いられている．これは，まず簡単な音響モデルと言語モデルを用いて N 位までの認識結果候補を選び，次に精度の高いモデルを用いてそれらの候補の順位を再評価することによって，認識性能を向上させる方法である．

新しい話者認識方法として，最近提案したテキスト指定型話者認識の構成を図2に示す．テキスト指定型とは，録音した音声によって装置がだまされるのを防ぐため，認識の度に新しい言葉を装置の側から指定する方法である．このため，その人のあらゆる文章音声がモデル化できるように，あらかじめ本人が発声した少数の文章音声から，すべての音素のモデルを学習する技術がキーになっている．近年の話者認識に共通して，音声認識と同様に，ケプストラムを特徴パラメータとするHMMを用いており，話者照合の判定のしきい値を安定に認定するために，発声時期や発声内容の変化によるゆう度値の変化を，事後確率の概念によって正規化している．

3. これまでの音声および話者認識技術の変遷

3.1 特徴分析

音声認識の研究は，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[1]．初期の音声認識では，一人の話者が区切って発声した一けた数字，単音節などを認識対象としていたので，音声の特徴分析と，蓄積された特徴との比較法が研究の中心であった．特徴分析としては帯域フィルタバンクを用いるのが主流であったが，音韻性に密接な関係があるスペクトル包絡を精度良く表

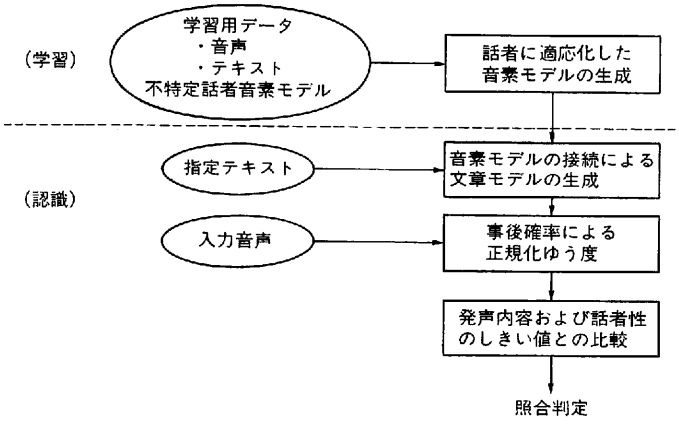


図2　テキスト指定型話者認識の構成　　認識の度に新しい言葉を装置の側から指定する．このため，その人のあらゆる文章音声がモデル化できるように，あらかじめ本人が発声した少数の文章音声から，すべての音素のモデルを学習しておく．

現するのが難しいという問題点があった.

そこで1960年代の後半に板倉ら（NTT）（敬称略）によって提案されたのが，最ゆうスペクトル推定法である．これはスペクトル包絡を全極型のモデルで表現し，最ゆう法の意味で最適なモデルパラメータを，比較的少ない計算量で，安定に抽出できる優れた性質をもっている．同じことが，時間領域では波形の線形予測モデルとして定式化できるので，線形予測分析（LPC：Linear Predictive Coding）法ともよばれる．

同じ1960年代に，スペクトル包絡を表現する優れた方法として，ケプストラム法が提案されていた．ケプストラムは，対数スペクトルの逆フーリエ変換として定義され，畳込みの関係にある現象を分離して表現できる性質があり，音声のスペクトル包絡とピッチ（基本周波数）を分離抽出できる方法として注目されていた．この性質は，線形予測分析法に基づくピッチ抽出法，すなわち変形相関法の提案にも影響を与えたと考えられる．

線形予測分析法は，その後，音声の分析合成の優れた技術として，PARCOR（Partial Auto-Correlation）法，LSP（Line Spectrum Pair）法などへ発展し，現在の音声符号化でも基本的技術として用いられている．この華々しい成功は音声認識にも影響し，1970年代から，線形予測分析法に基づく音響モデルが，専ら用いられるようになった．

その中にあって，筆者は，1970年に話者認識の研究を始めた．話者認識は，音声認識よりもいわば一段階細かい情報を扱うものであり，筆者は統計的パターン認識理論にのっとった方法が必要と考えた．この観点から考えると，ケプストラムには線形予測分析から導かれるパラメータよりも優れた性質がある．

（1） パラメータによって表現されるスペクトル包絡が，滑らかで安定した性質を有している．

（2） パラメータの単純なユークリッド距離が，対数スペクトル包絡の距離に一致する．

（3） パラメータが直交化されており，パラメータ間の相関が少ない．

（4） パラメータの分布が，経験的に正規分布に近い．

このため筆者はケプストラムを用いることを選んだ．ケプストラムの抽出に必要な計算量は，線形予測分析よりもやや大きかったが，FFT（高速フーリエ変換）を用いれば，問題はなかった．1970年代の半ばには，ベル研のAtalによって，線形予測係数から全極型スペクトル包絡のケプストラムを直接的に求める方法が提案され，計算量の問題は完全に解消された．

更に，1970年代の終りには，聴覚に重要な音声スペクトルの動的性質を話者認識に用いる方法として，ケプストラムの時系列（50～90 ms程度の長さ）を，10 ms程度の通常の音声の分析フレームごとに多項式で展開し，その展開係数を特徴パラメータとして，元のケプストラムと組み合わせて用いる方法を提案した（これらの展開係数は，その後，Δケプストラムおよび$\Delta\Delta$ケプストラムと名付けられた）．その後1980年代の半ばに，この方法を一般の音声認識に用いることを提案した．この方法は，音声の個人差，雑音，ひずみなどに強く，音声認識の精度を大きく改善する効果があることが，多くの実験によって示された．その結果，それまでの線形予測分析法に基づく方法に代って，世界中のほとんどの音声認識システムでこの方法が用いられるようになった．これだけ広く用いられるようになった理由は，統計的な方法が音声認識の中心的方法となったため，上記のケプストラムのもつ優れた性質が，より生きるようになったことが挙げられる．

3.2 DPマッチングとHMM

1970年代の初めの，日本における重要な研究成果の一つに，迫江ら（NEC）による，動的計画法（DP：Dynamic Programming）を用いた時間軸の整合法（DPマッチング法）がある．これは音声の時間軸の伸縮に対処しながら，二つのパターンの類似度（距離）を計算する極めて効率的な方法である．同じころにソビエト連

邦（現ウクライナ）でも研究されていたことが，後にわかっている．この音声の時間軸伸縮の正規化に動的計画法を用いる方法は，二段DPマッチング法やその変形として，連続単語音声認識における単語列のマッチングにも用いられ，現在のHMMによる音声認識でも用いられている．

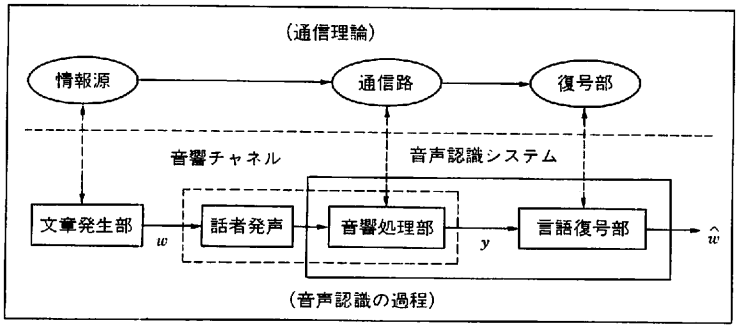
1980年代は，それまでの音声の時間パターンを直接整合させる方法から，HMMを代表とする統計的モデル化に基づく方法への移行によって特徴づけることができる．HMMによる方法は，IBMなど少数の研究所ではそれ以前からよく知られていたが，一般的には，1980年代の半ばにその具体的方法と理論がベル研の研究者などによって広く発表されるようになって，初めて世界の研究機関で用いられるようになった．

3.3　言語モデル

統計的言語モデルのさきがけは，1950年代に英国のUniversity Collegeで行われた音素認識システムにあり，この中で始めて英語の音素連鎖に関する統計的情報が用いられた．しかしその後20年以上にわたって，この方法が音声認識で積極的に用いられることはなく，生成規則による方法が専ら用いられていた．

1980年代になって，上述のように音響処理にHMMが広く用いられるようになるのと歩調を合わせて，言語モデルでも統計的方法が中心的に用いられるようになった．この基本になっている音声の発声と認識のモデルは，図3に示すとおりで，通信理論と対応づけて定式化されている．ここでは，話者は情報源に対応する文章wを，その話者の癖に従って発声し，音声認識システムは，その音声から音響処理（分析）によって特徴パラメータyを得る．このモデルでは，話者と音響処理を一つにまとめて音響チャネルと考え，通信理論において雑音が加わる通信路と対応させている．言語復号部はyをもとに，事後確率$p(w|y)$が最大となるwを求めることを目的とするが，これは直接的に計算することができないので，ベイズの規則に従って変形し，同時確率$p(w, y)=p(y|w)p(w)$を最大化する．この内，$p(y|w)$はHMMによって与えられ，$p(w)$は統計的言語モデルによって与えられる．

1980年代に，統計的言語モデルが特に米国において広く用いられるようになった直接的原因は，ARPA（Advanced Research Projects Agency）による，1,000～20,000の語彙を対象とした連続音声認識プロジェクトによって，ばく大な量の言語および音声データベースが構築されるようになったことにある．認識結果の評価法を明確にして，結果に関するコンテストを毎年行うというし烈な競争を研究に導入したこ



$$p(\hat{w}|y)=\max_{w}\frac{p(y|w)\ p(w)}{p(y)}$$

図3　現在の音声認識システムの基本になっている発声と認識のモデル　話者は情報源に対応する文章wを発声し，音声認識システムは，その音声から音響処理（分析）によって特徴パラメータyを得る．言語復号部はyをもとに，事後確率$p(w|y)$が最大となるwを求める．

とと合わせて，世界の音声認識の技術レベルの向上に，このプロジェクトの果たした役割は極めて大きい．残念ながら，我が国においては，このような共通の言語データベースを構築する動きが機能していないため，このような研究は大きく遅れている．

3.4 ニューラルネット

1980年代のニューラルネットブームの中で，音声認識の分野においてもニューラルネットを用いる試みが行われ，限定されたタスクにおいては，高い認識性能を示して注目されたこともあった．しかし，結局HMMのような統計的方法を明確に凌駕（りょうが）することはできなかった．また，連続音声認識のように，音響モデルと言語モデルを複雑に組み合わせなければならない場面では，非線形な処理の結果を総合的なスコアへ積み上げたり，文章全体としての認識性能が高まるように，パラメータの最適化を図るのが極めて難しい．このため，音声認識におけるニューラルネットの研究は，最近ではすっかり下火になってしまった．

但し，ニューラルネットの一種である学習ベクトル量子化(LVQ：Learning Vector Quantization)の理論的発展として生まれた，誤識別最小化学習（MCE/GPD）法は，今後の学習法の一つの基本的考え方として，現在注目されている．

4. 音声および話者認識技術の21世紀への展望

音声認識技術は，最近米国を中心に具体的な応用が広く開拓されつつあり，今後，マルチメディア環境の中で，他のメディアと組み合わされながら，種々の形で使われるようになると期待されている．主な用途としては，データベース（情報）検索・案内，自動予約・注文，ディクテーション，電子秘書，身体障害者の補助機器などがある．日本では，今から10年ほど前に，関西にATR（(株)国際電気通信基礎技術研究所）の音声と言語に関連する研究グループが発足し，自動翻訳電話を目標とした研究が活発に行われた．第1次のプロジェクトの終了時には，日米独をつないだデモも行われたが，実際に役に立つシステムを構築するためには，越えなければならない壁が多数存在している．

話者認識技術は，上記の情報サービス，バンキングサービス，買い物サービス，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，音声を本人確認に用いる手段として用いられるようになると期待される．

音声認識全般として，現在の統計的方法をベースに，実際の大量の音声データに基づいて，話し言葉の言語モデル（規則）を構築すること，多数の話者の音声データに基づいて個人差のモデルを構築し，それによる話者の声への適応化アルゴリズムを開拓すること，種々の雑音やひずみに自動的に適応できる方法を確立することなどが，重要な技術的課題となろう．

文　献

(1) 古井貞熙，"音声認識の過去・現在・未来，" NTT R & D, vol.43, no.10, pp.1055-1060, Oct.1994.

古井　貞熙（ふるい　さだおき）（正員）

昭43東大・工・計数卒．昭45同大学院修士課程了．同年日本電信電話公社（現NTT）電気通信研究所入所．以後，同所において，音声認識，話者認識，音声知覚などの研究に従事．昭53～54ベル研究所客員研究員．現在，NTTヒューマンインタフェース研究所古井特別研究室長．東工大客員教授．工博．昭49年度米沢賞，昭62年度，平4年度論文賞，平元年度著述賞，平元科学技術庁長官賞，IEEE ASSP Society Senior Awardなど受賞．著書「ディジタル音声処理」，「Digital Speech Processing, Synthesis, and Recognition」，「音響・音声工学」，編著「Advances in Speech Signal Processing」など．IEEE Fellow.