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)	ー音響学における20世紀の成果と21世紀に残された課題ー 分野別の流れー音声分野
Title	Overview in Each Research Field: Speech
著者(和文)	板橋秀一, 粕谷英樹, 北脇信彦, 中川聖一, 匂坂芳典, 古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会誌, Vol. 57, No. 1, pp. 11-20
Journal/Book name	, Vol. 57, No. 1, pp. 11-20
発行日 / Issue date	2001, 1

Powered by T2R2 (Science Tokyo Research Repository)

日本音響学会誌 57 巻 1 号 (2001), pp. 11-20

特集—音響学における 20 世紀の成果と 21 世紀に残された課題—

第 1 部—分野別の流れ— 音声分野*

板橋秀一 (筑波大学)・粕谷英樹 (宇都宮大学)・北脇信彦 (筑波大学)
中川聖一 (豊橋技術科学大学)・匂坂芳典 (ATR)・古井貞熙 (東京工業大学)**

1. 音声研究の始まり

本特集号は音声研究における 20 世紀の成果と 21 世紀に残された課題を 4 分野 (生成・分析・認識・合成) に分けてそれぞれの専門家に述べていただき, 最後に今後の展開を総括しようとするものである。

現在の音声研究の源は H. Dudley の Voder/Vocoder (1939) にさかのぼることができるが, 更にそれ以前の 1937 年に小幡・小林両氏によりピッチ抽出回路の論文が JASA に発表されていたことから, 日本の音声研究は当時かなり先進的であったといえよう[1]。

各研究の詳細な記述は他の編を参照いただくとして, ここでは筆者が音声研究の道に入った 1963 年以降の日本を中心とした音声研究の状況を概観してみたい。その翌年の 1964 年度以来, 文部省科研費による音声に関する総合研究が, 3 年ごとに代表者が交代して実施され 15 年間続いた。これには工学のみならず, 言語学, 音声学, 生理学の研究者も加わり, 学際的な集まりであって, その後の音声研究の進展に大きな役割を果たした[1]。当時大学院生であった筆者は大きな刺激を受けた。

この頃に利用できた参考書は, 千葉・梶山 (**第 2 部 "The Vowel: Its Nature and Structure" の出版参照**, p. 64) や G. Fant の本, Potter 他による "Visible Speech" (1947) 等であり, J.L. Flanagan「音声分析・合成・知覚」(1965) が刊行されたばかりであった。日本語では勝木・三浦「聴覚と音声」(1966) がほとんど唯一のものであった。その後 1970 年代後半からは日本語の参考書が次第に出版されるようになった。

当時, 研究のための分析にはサウンドスペクトログラフ (**第 2 部サウンド・スペクトログラム参照**, p. 65) を使い, 認識処理のための分析にはハードウェアのフィルタバンクを使うのが普通であった。その後, 計算機の発達により, FFT (**第 2 部 FFT を用いれば参照**, p. 85) やケプストラムが次第に用いられるようになった。

世界では MIT やスウェーデンの王立工科大, ベル研等が音声研究の中心であり, 日本から多くの方が留学 (**第 2 部日本の音声生成研究の発展と欧米での研究参照**, p. 63) している。音声分析・知覚の研究が盛んに行われ, 音声合成では, 声道模擬型, ホルマント型等の方式が研究の中心であった。更に R. Reddy を中心とする (D)ARPA プロジェクト (**第 2 部 DARPA 音声プロジェクトと日本の音声認識研究参照**, p. 70) では, 計算機処理による音声認識・理解に関する研究が大規模に進められ, その後の研究の方向に大きな影響を与えた。規則音声合成方式では, アメリカのハスキンス研究所やイギリス等で研究が進められ, MIT の MITALK が一時期を画した。

日本に戻ると, 音声生成系では, 石坂先生による声帯の 2 質量モデル (**第 2 部調音モデル参照**, p. 64), 韻律生成では藤崎モデル等が広く知られ用いられている。音声生成機構の観測 (**第 2 部 X 線撮影と音声研究参照**, p. 62) では, 東京大学の X 線マイクロビーム法や熊本大の磁気センサ法等の研究が進められた (**第 2 部調音モデル参照**, p. 64)。NTT の板倉氏 (現, 名大教授) や B.S. Atal により LPC が提唱され, その発展形としての PARCOR, LSP は音声分析・合成全般に大きな影響を与え (**第 2 部線形予測分析 LPC の発明→ PARCOR → LSP へ参照**, p. 66), 更に低ビットレート伝送方式の実現をもたらした (**第 2 部携帯電話・PHS 実現の鍵 波形符号化参照**, p. 67)。また, 東工大では一般化ケプストラムモデルの研究が始められた。一方, 電総研では声道断面積形抽出の研究が進められ, 後には連続 DP 法が提唱された (**第 2 部 DP *vs.* HMM 参照**, p. 68)。音声認識では, 日本電気の迫江氏 (現, 九大教授) や Velichko, Zagoruiko による DP 法の適用が一つの契機となって単語音声認識の実用化につながった。

* Overview in each research field: Speech.
** Shuichi Itahashi, Hideki Kasuya, Nobuhiko Kitawaki, Seiichi Nakagawa, Yoshinori Sagisaka and Sadaoki Furui.

1980 年代に入ると HMM の進展と共に音声データベースの必要性が認識されるようになり，東京で行われた ICASSP 86 では電子協音声データベースの報告が行われた（**第 2 部 DP *vs.* HMM 参照**，p. 68）。現在ではテキスト音声合成には PSOLA やコーパスベースの手法が用いられるようになり，音声データベースの重要性が更に増している（**第 2 部コーパスベース音声合成参照**，p. 71）。

21 世紀における音声研究の一層の発展を期待して本稿を結びたい。

1) 藤崎博也，"日本の音声研究と共に歩んだ 30 年の回顧と将来への期待，" 音響学会誌 **45**, 906-909 (1989).

（板橋秀一）

2. 音声生成

人間の音声器官と聴覚器官はどちらも魚の鰓に起源を持つといわれる[1]。脊椎動物の進化の歴史のなかで，音声器官と聴覚器官は哺乳類化の過程で分離して進化し，現在ヒトに見られる聴覚器官の構造をほぼ完成させた。一方，音声器官はいわゆるヒト化の段階で大きな変貌を遂げた。ヒト化の原点は二足歩行にあるといわれる。自由になった手の使用が脳の拡大をうながし，それが頭蓋底を屈曲させ，更に喉頭軟骨が舌骨と分離して下降することによって，咽頭腔が広がり，舌の移動を大きくしかも喉頭の発声機構とはほぼ独立に制御できるようになった。喉頭における発声（音源）と声道における調音（フィルタ）をそれぞれほぼ独立に制御できるのはヒトだけである。このように音声言語の生成の仕組みを発声と調音という考えのもとに，科学的かつ体系的に解明したのは，当時東京外国語学校（現東京外国語大学）の音声学研究室に勤務していた千葉勤と梶山正登が最初である。

千葉・梶山は 1934 年に理論的並びに実験的研究を開始し，その後 5 年間の研究の成果をまとめて発表したのが，音声生成の音響理論の端緒となった The Vowel : Its Nature and Structure[2] である（**第 2 部 "The Vowel : Its Nature and Structure" の出版参照**，p. 64）。この研究の特色はなんといっても，自らの工夫によるものも含めて X 線撮影装置など当時最先端の実験装置を駆使して日本語母音発音時の生理学的データを収集し，音響理論と電気回路理論に基づいて数値計算を行って声道における共鳴（物理現象）が声道断面積関数によって決定されることを明らかにしたことである。両氏は更に聴覚の内耳における低分解能の周波数分析機構の考えを適用して，共鳴が母音の音質（知覚現象）を決める要因になっているという説明も行っている。千葉・梶山の業績は 1940 年代の後半から G. Fant (Royal Institute of Technology, RIT) へと引き継がれた。Fant は子音の生成機構にまで適用範囲を広げながら音響理論を一層精緻化し，いわゆる音声生成のソース・フィルタ理論（**第 2 部ソース・フィルタ理論参照**，p. 65）として体系化した。1960 年にその集大成として "Acoustic Theory of Speech Production[3]" を出版した。Fant も X 線写真によって測定した声道形状を基礎資料として用いたが，そのときの言語はロシア語であった。これは R. Jacobson (Harvard University) 及び M. Halle (MIT) との共同研究によるものであるが，この共同研究は後の音声学・音韻論に大きな影響を与えた弁別素性理論へと発展することになる[4]。子音の音響理論では K.N. Stevens (MIT)[5] や藤村靖（電気通信大学）らも大きな貢献をした（**第 2 部日本の音声生成研究の発展と欧米での研究参照**，p. 63）。

これまで述べた研究の過程は，主として声道形状と音声のスペクトル特性の関係（フィルタの性質）を論ずるものであったが，音声生成機構のもう一方の要素であるソース（音源）となる声帯振動がどのようにして持続できるかを説明する理論が現れるまでには，Fant の出版から 8 年待たなければならなかった。石坂謙三（電気通信大学）・松平正寿（玉川大学）は，1968 年に東京で開催された第 6 回国際音響学会議において，声帯振動の 2 質量モデルという画期的な論文を発表した。同じ会議で J. Flanagan (Bell Telephone Lab., BTL) は 1 質量モデルによるコンピュータシミュレーションを発表した。しかし 1 質量モデルの不備については石坂・松平が既に指摘していた。会議終了後まもなく Flanagan は石坂を BTL に招聘して，声帯振動の 2 質量モデルと声道調音モデルを一体化した有声音生成機構のシミュレーションを開始し，二人の共同研究が長期間にわたって続くことになる（**第 2 部日本の音声生成研究の発展と欧米での研究参照**，p. 63）。記念的論文になったのが，Ishizaka と Flanagan が共著であったため[6]，2 質量モデルが Ishizaka-Flanagan モデルと呼ばれるが，持続振動の本質的な部分は既に石坂・松平が解決していた。

それまで声帯振動の定説になっていた Van Den Berg の粘弾性空気力学理論の欠陥を理論的実験的に確認し，広戸幾一郎（久留米大学）が注目していた声帯粘膜波状運動に着想を得て，一番簡単な 2 自由度モデルを構築して，声道負荷にほとんど依存しな

い自励声帯振動機構を明らかにした。声帯の下唇が上唇より進み位相で振動することが本質的であり，それによって気流がジェット状になって上唇で声帯表面から離れ，気流から声帯へのエネルギーの実効的な伝達が行われることを明らかにした。その後I. Titze (University of Iowa) は声帯組織についての平野実（久留米大学）のボディー・カバー理論に基づいて，多自由度系へ拡張すると共に，可能な声帯振動モードをリボンモデルによって整理した[7]。このように，ソース・フィルタ理論（**第2部ソース・フィルタ理論参照**，p. 65）の大切な部分で日本の研究者が大きな貢献をしていることは特筆すべきことである。

これまでは，どちらかと言えば定常的な音声生成の音響理論の発展について述べた。しかし通常の発話は定常的な音声単位を時系列的に接続して構成されているわけでわない。発話のなかの音の性質は前後の音の影響を受けて変化するが，それは生理的・音響的な面を含めていろいろなレベルで観測される。このような適応現象のなかで，特に顕著なものが調音結合である。例えば，牙/kiba/と木場/koba/の第1音節の子音は同じ音素/k/として表記されるが，前者の［k］では舌が口蓋に接触する点が後者の［k］よりも前寄りであり，その違いがフォルマント周波数遷移にも現れる。このような子音の調音点の移動は後続の母音の調音点に同化するかたちで結合していると解釈される。このような調音結合は音声生成のどの段階で起こっているのか，また，調音結合によって変化した音響刺激から人間はどのように言語音を知覚しているのか，あるいは言語音を決める不変量はなにか，について1960年代以降，Haskins研究所，王立工科大学(RIT)，MITを中心に積極的に研究された。そのなかでもHaskins研究所のLiberman[8]らの，知覚モデルと対になる音声生成のコード化モデルが著名である。中枢神経の段階で複数の音素が並列的に処理され，筋への神経指令が構築されて筋の運動指令となり，それによって筋が活動し，音声器官各部固有の運動として実現し，音声が生成される，というものである。このようなモデルを実証するには，音声器官の運動や筋活動などの観測が欠かせない。観測技術の発展には，東京大学医学部音声言語医学研究施設（**第2部X線撮影と音声研究参照**，p. 62）が大きな貢献をした。X線被曝量を最小限に押さえながら，舌などの調音器官の運動を観測するためのX線マイクロビーム撮影装置，声帯の詳細な振動を観測するためのディジタル高速度撮影装置などの開発のほか，発話中の筋電図の計測では多くの研究業績を残した。また，同施設の廣瀬肇らとHaskins研究所との共同研究も大きな成果を上げた（**第2部日本の音声生成研究の発展と欧米での研究参照**，p. 63）。

前述のような複雑な階層構造を科学的に理解するためには，調音器官の運動と音声波を繋ぐ調音モデルの役割も大きい。この研究は音声合成の研究とも関連して，BTL，MIT，RIT，CNRS (France) のほか日本でも電子技術総合研究所，東北大学，早稲田大学，NTT，ATRなどで精力的に研究されてきた。1970年ごろに比企静雄ら（東北大学）は神経指令から音声波までをつなぐ階層的計算モデルの研究を行ったが，その基礎となる生理学的データが十分得られなかったために，研究半ばでプロジェクトを終了した。一方，Ishizaka-Flanaganは前述のモデル[6]を用いて音声波から直接生理学的モデルパラメータを推定する研究を行った。最近のMRIや多チャンネル磁気センサなどの観測技術の進歩を背景に，ATRやNTTで調音モデルの研究が進められているが，問題の複雑さ・難しさを考えれば，成果を急ぐよりは，積み重ねがきく研究を期待したい。

ピッチ，強度，継続時間，声質などは言語のみならずパラ言語や非言語（個人性）などの特徴を担う要因として極めて重要である。ピッチ軌跡を生成するための数理モデルとして，藤崎博也ら（東京大学）が提案したいわゆる"Fujisaki model"は広く受け入れられ，いろいろな言語に適用されている（**第2部日本の音声生成研究の発展と欧米での研究参照**，p. 63）。ピッチなどの超分節的特徴に関わる喉頭の生理学的制御機構についてはまだ分からないことが多く，今後の発展が望まれる。

日常の豊な会話は，言語情報だけでなくパラ言語，非言語情報によって成立している。これらの情報が神経生理的レベルから音声波まで互いに干渉し合いながらどのようにコード化されるかを明らかにするための研究は，21世紀に積み残された課題である。

1) 本多清志，言語の科学—第2巻（岩波書店，東京，1998）.
2) T. Chiba and M. Kajiyama, *The Vowel : Its Nature and Structure* (Tokyo-Kaiseikan Pub. Co., Ltd., Tokyo, 1941).
3) G. Fant, *Acoustic Theory of Speech Production* (Mouton & Co., The Hague, 1960).
4) R. Jacobson, G. Fant and M. Halle, "Preliminaries to speech analysis : The distinctive features and their correlates," MIT Acoust. Lab., Tech. Rep. No. 13 (1952).
5) K.N. Stevens, *Acoustic Phonetics* (The MIT Press, Cambridge, Mass., 1998).
6) K. Ishizaka and J. Flanagan, "Synthesis of voiced

sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J. 51, 1233-1268 (1972).
7) I. Titze, *Principles of Voice Production*, (Prentice Hall, Englewood Cliffs, 1994).
8) A.M. Liberman, F.S. Cooper, D.P. Shankweiler and M. Studdert-Kennedy, "Perception of the speech code," Psychol. Rev. 74, 431-461 (1967).

(粕谷英樹)

3. 音声分析と符号化

3.1 はじめに

Fant の音響的音声生成理論によれば，肺からの空気流及びそれによる声帯振動が音源となって，声道における音響的共鳴を起こさせ，意図した音声波が発生される（**第 2 部ソース・フィルタ理論参照**，p. 65）[1]。このとき音源の生成と声道の共鳴による調音とは分離され，電気的等価回路で表された各々の過程は，相互作用を伴わずに接続される。

音声器官によって生成され空気中に放射された音声波から，種々の物理的特徴あるいは統計的特徴を明らかにすることが音声分析の目的である。音声分析の歴史は，1946 年に開発されたソナグラフ (sound spectrograph の略) と呼ばれる電気回路から始まった[2]。ソナグラフは帯域フィルタ群によってサウンドスペクトログラム（音声周波数スペクトルを時間的に記録したもの）を抽出する装置である（**第 2 部サウンド・スペクトログラム参照**，p. 65）。

帯域フィルタ群によって分析された音声スペクトルを特徴パラメータとする低ビットレート符号化方式が，1939 年に提案されたチャネルボコーダである[3]。チャネルボコーダは分析合成系による音声符号化の草分けとなった。

3.2 音声のスペクトル分析

電気回路による音声分析の時代から，コンピュータ処理の時代を切り開いたのが，1967 年に提案された最尤スペクトル推定法であった[4]。これはスペクトルの構造にある形のモデルを設定し，そのモデルを尤度比最大という条件のもとで推定するものであった。翌 1968 年には，音声波形の線形予測分析法が提案された[5]。両者はスペクトル次元と波形次元との双対な関係にあることが分かり，線形予測符号化法 (LPC : linear predictive coding) として体系化された（**第 2 部線形予測分析 LPC の発明→ PARCOR → LSP へ参照**，p. 66）。

LPC では，時系列信号波形を線形予測モデルにあてはめて，信号の持つ統計的特徴を抽出する。適応予測モデルは，全極型 AR (Auto Regressive，自己回帰)，全零型 MA (Moving Average，移動平均)，極零型 ARMA 予測に分類でき，予測係数の更新方法は前方適応予測と後方適応予測に分類できる。最もよく用いられる AR 予測では，信号波形を過去のサンプル値の加重平均で表現し，その線形予測誤差の二乗平均を最少にするという条件で，線形予測係数を求める。具体的な解法には，自己相関法 (auto-correlation method) と共分散法 (covariance method) がある。

LPC のように，音声波の発生に関してモデルを設定して分析する方法をパラメトリック分析 (parametric analysis)，特定のモデルを設定せずに行う方法をノンパラメトリック分析 (non-parametric analysis) という。

ノンパラメトリック分析にはフーリエ分析 (Fourier analysis)，帯域フィルタ分析 (band-pass filter analysis)，自己相関分析 (auto-correlation analysis)，ケプストラム分析 (cepstrum analysis) などがある。

このような音声スペクトルを表現するパラメータについて多くの研究が進められ，音声符号化だけでなく，音声認識，音声合成，その他様々な信号処理用途に利用されている。

3.3 分析合成系

音源と調音の分離構造を持つ符号化法を分析合成系，分離構造を持たない方式を波形符号化法という。調音パラメータとして音声スペクトル包絡パラメータを用いる分析合成系をスペクトル符号化法ともいう。

線形予測係数をスペクトルパラメータとする最尤ボコーダは，それまでの電気回路型分析合成系に比べて，音質の点で格段に優れていた。しかし，低ビットレート化するために線形予測係数を量子化すると，合成器の安定性が保証されないという欠点があった。この問題を解決した方式が 1969 年の PARCOR (PARtial auto-CORrelation，偏自己相関)[6] 及び 1975 年の LSP (Line Spectrum Pair，線スペクトル対)[7] である（**第 2 部線形予測分析 LPC の発明→ PARCOR → LSP へ参照**，p. 66）。

PARCOR は，線形予測分析において，音声波形の二つの標本値間にはさまれた複数個の標本値から，前方適応予測した予測誤差と後方適応予測した予測誤差との相関係数として定義される。音声波形の相関性を逐次取り除くことによって，音声波形から直接求められ，合成器の安定性も保証される。

PARCOR は Kelly の音響管モデルの反射係数に相当するが，パラメータと音声スペクトル構造が直接対応しないので扱いにくいきらいがあった。Kelly の音響管モデルにおいて，唇端は開放とし，声門を開

放と閉塞の2条件で求めた線スペクトル対がLSPである。LSPは線形予測係数と相互変換が可能で，スペクトル構造とも直接対応する。LSPパラメータは量子化特性と補間特性にすぐれているため，ITU，MPEG，PDCなどの標準化機構における低ビットレート標準音声符号化方式のほとんどすべてで用いられている。

3.4　残差信号の能率的符号化法

音声波形の線形予測分析による誤差波形を残差信号という。残差信号には，音源情報を表すスペクトル微細構造成分が含まれている。分析合成系では，有声音源をパルス発生器で，無声音源を白色ランダム雑音発生器で構成し，その制御パラメータを残差信号から得ていた。必要とするビット数はわずかであるため，極低ビットレート符号化が可能であった。しかし，あまりにも情報量を少なくしていることから電話品質に及ばないこと，ロバスト性に欠けることなどの問題があり，残差信号をいかに効率よく低ビットで表現できるかが重要な課題であった。

1985年に発表されたCELP（Code Excited Linear Prediction，符号励振線形予測）が，これに対する一つの答えであった（**第2部携帯電話・PHS実現の鍵　波形符号化参照**，p. 67）[8]。CELPでは，音声の音韻情報をスペクトル符号化し，音源情報を波形符号化するハイブリッド方式を採っている。音源情報は残差波形をベクトル量子化して符号帳（code book）を作成しておき，入力と符号帳とのマッチングを行い，最適符号帳につけられた番号のみを伝送する。代表的なベクトル量子化による符号帳作成アルゴリズムには，1980年に発表されたLBG（Linde-Buzo-Gray）法がある[9]。

20世紀末に爆発的に普及した携帯電話やインターネットで用いられた音声符号化方式は，LSPをスペクトルパラメータとし，CELP型の構造を持つものであった（**第2部携帯電話・PHS実現の鍵　波形符号化参照**，p. 67）。代表的な方式として，ITU勧告G. 729 CS-ACELP，世界標準携帯電話（IMT-2000）のW-CDMA，我が国携帯電話のPSI-CELPなどがあげられる。

3.5　楽音の分析合成

音声分析技術の進展は，楽音の符号化にも影響を及ぼした。CDを楽音ディジタル化の原点とし，音声符号化で開発された種々の分析技術が適用され，1992年にMPEG-1，1996年にMPEG-2，1999年にMPEG-4が標準化された。注目すべきは，1987年のMDCT（Modified Discrete Cosine Transform，変形離散余弦変換）の利用と1977年のQMF（Quadrature Mirror Filter，直交鏡像フィルタ）の利用である。

MPEG-1は32帯域のサブバンド符号化，MPEG-2はそのマルチチャネル化，MPEG-4は低レート化を特徴としている（**第2部オーディオの高能率符号化参照**，p. 79）。これらにより，楽音メディアのコンパクトなパッケージ（MD）への記録やインターネットによる音楽配信（MP3やTwinVQ）が盛んになった。

3.6　今後の展望

音声符号化の理論は1948年のShannonの論文（A mathematical theory of communication）に始まる。Shannonの理論では，最適な情報源符号化と伝送路符号化を組み合わせれば最適なシステムが構築できるとされている。最近では，情報源符号化と伝送路符号化とが融合した方式が開発されており，この50年間の符号化技術の進歩はShannonの予想を上回るものだったかも知れない。

これからは，WI（Waveform Interpolation），MBE（Multi-Band Excitation vocoder），HVXC（Harmonic Vector eXcitation Coding）に見られるように，更に音源を工夫して低レート・高品質化が研究される。また，音声，楽音，映像を総合的に処理するマルチメディア符号化の研究が期待される。

1) G. Fant, *Acoustic Theory of Speech Production* (Mouton & Co.'s-Gravenhage, The Hague. 1960).
2) W. Koenig, *et al.*, "The sound spectrograph," J. Acoust. Soc. Am. **18**, 19-49 (1946).
3) W.H. Dudley, "The vocoder," Bell Labs. Rec. **18**, 122 (1939).
4) F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," 6th ICA, C-5-5 (1968).
5) B.S. Atal and M.R. Schroeder, "Adaptive predictive coding of speech signals," Bell Syst. Tech. J. **49**, 1973-1986 (1970).
6) 板倉文忠，齋藤收三，"偏自己相関係数による音声分析合成系，" 音響学会音声研資 (1969).
7) 板倉文忠，"線形予測係数の線スペクトル表現，" 音響学会音声研資，S 75-24 (1975).
8) M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP) : high-quality speech at very low bit rates," IEEE Int. Conf. Acoust. Speech Signal Process., 937-940 (1985).
9) Y. Linde, A. Buzo and R.M. Gray, "An algorithm for vector quantization design," IEEE Trans. Commun. **COM-28**, 84-95 (1980).

（北脇信彦）

4. 音声認識

4.1　研究の流れ

人間の発した音声を機械で正しく認識させるという音声認識の研究は，古くは1950年前後から行われ

てきた（Davis らの数字認識など）。我が国でもハードウェアによる認識装置が 1950 年代後半から 1960 年代の初めに開発された（鈴木・中田の母音認識や坂井・堂下の音声タイプライタなど）。当時としては画期的な研究であったが，すべてハードウェアで構成しなければならないという制約から研究はあまり進まなかった。1960 年代に入って，音声波形を直接計算機に取り込んだり，アナログフィルタバンクによる音声の分析後，計算機に取り込むことができるようになってから，音声研究は共通基盤を見出し，学問らしい研究となった。1960 年代後半には，音声の生成モデルとして自己回帰モデルを採用した線形予測分析法（**第 2 部線形予測分析 LPC の発明→PARCOR → LSP へ参照**，p. 66）や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（Vintsyuk など，**第 2 部 DP *vs.* HMM 参照**，p. 68）が考案された。

この二つの手法が，1970 年代の音声認識技術の主流となり，線形予測分析で得られるパラメータからのスペクトル同士の種々の比較法（スペクトル距離尺度）や，連続して単語を発声した音声を認識する連続音声認識のための DP マッチング法などが次々と考案され，ディジタルシグナルプロセッサを用いることによって研究室段階で，種々の音声認識のデモが行えるようになった。なかでも，追江・千葉による DP マッチング，追江による 2 段 DP 法による連続単語認識，板倉によるスペクトル距離尺度と DP マッチングの併用による孤立単語の認識は，1970 年代前半を代表する研究成果である。

1971 年から 5 年間実施されたアメリカの APRA（国防省高等研究計画局）による「音声理解プロジェクト」の貢献は非常に大きく，音声認識における言語知識の役割の重要性を示した（**第 2 部 DARPA 音声プロジェクトと日本の音声認識研究参照**，p. 70）。今日の原点となる発音辞書，木探索法（ビームサーチなど），韻律情報の利用，文脈自由文法の効率的利用法，ネットワークモデル・階層モデル・黒板モデルなどのシステムアキーテクチャ，などあらゆるアプローチ・問題がこの時代に検討された。このプロジェクトの成果は知識分散型の HEARSAY のような人工知能的手法よりも知識集約型の HARPY のようなパターンマッチング手法の優位性を示したことであろう。筆者らも音声理解システム LITHAN の開発を行い（1973〜1977 年），教師なし MAP 学習（1973 年），並列木探索法と称したビームサーチ法（1974 年），直接マッチング法と称したワードスポッティング法（1975 年，連続 DP 法と同一），Earley 法に基づく文脈自由文法制御の認識法（1975 年）などを一早く提案した。CMU の Baker や IBM ワトソン研究所の Jelinek らによる情報理論に基づく連続音声認識のアプローチ（隠れマルコフモデル（HMM）や n-gram，**第 2 部音声認識の情報理論・統計的アプローチによる定式化参照**，p. 69）の研究がこの時期になされ，今日の音声認識の実用化時代の基礎が築かれた。

1980 年代になって，大量の音声データが手軽に扱えるようになり，音声を統計的手法で認識するというパターン認識のオーソドックスな方法が主流になってきた。その枠組みの中心になったのがベル研の Levinson や Rabiner らにより分かり易く紹介された HMM であった（**第 2 部 DP *vs.* HMM 参照**，p. 68）。HMM は，DP マッチングを包含しており，話者やコンテキストによる音声パターンの変動を確率・統計的にモデル化するものである。これにより，音声認識技術は飛躍的に向上し今日に至っている。これには，トライフォンなどのコンテキスト依存モデルの有効性とワードペアなどの言語モデルの有効性を実証した K.F. Lee の博士論文の貢献が大きい。また，古井によって提案された動的特徴を表す 1 次回帰係数（デルタパラメータ）は DP マッチングだけでなく HMM にも有効で，広く用いられている。この期間で，研究の牽引力になったのが，DARPA の「大語彙連続音声認識プロジェクト」（**第 2 部 DARPA 音声プロジェクトと日本の音声認識研究参照**，p. 70），ヨーロッパの「ESPRIT プロジェクト：音声と画像のための高度アルゴリズムとアーキテクチャ」，我が国では ATR における「翻訳電話プロジェクト」と文部省の重点研究「音声言語」である。

一方，1980 年代後半頃から人工ニューラルネットワークがブームになり，一時パターン認識の革新的手法としてもてはやされ音声認識への適用に期待が集ったが，従来の統計的パターン認識の枠組みからあまり出ず従来手法を越えるには至っていない。また，同じ頃，人工知能分野におけるエキスパートシステムの流行に影響され，古くから行われていた音素の特徴抽出に基づく音声認識法が知識工学的手法（音響・音声学の知識・ルール集合による推論に基づく手法）と姿をかえて復活したが，成功するには至らなかった。

1990 年代に入ってからは，音声研究の方向は一段と実用的研究に向ってスペクトルを広げてきた。つまり，大語彙連続音声，音声対話，マルチモーダルインタフェース，実環境での認識等であり，コンテキスト依存音響・音声モデル，大規模言語モデル，

モデルの自由パラメータの推定とパラメータ数の削減，意味抽出，非文法・未知語・不要語検出，雑音・残響等の扱いが中心となってきた。この方向は必然的ではあるが，ARPAの「音声言語理解プロジェクト（ATIS）」と我が国ではATRの組織的な研究と音声データベース，文部省の重点研究「音声対話」，日本音響学会研究用音声データベース（ASJ，JNAS）の果した役割は大きい（**第2部DARPA音声プロジェクトと日本の音声認識研究参照**，p.70）。1997年末の日本IBMの日本語ディクテーションシステムVia-Voiceの商品化，2000年2月からのNHKの音声認識技術を用いた字幕放送の実用化は我々音声認識研究者の努力の勝利と言えよう。

4.2　研究の成果と知見

最近の四半世紀の研究で明らかになったことは，工学的には，時系列パターンを扱う確率モデルを定義し，できるだけ正確にモデルパラメータを推定し（最尤推定やMAP学習の他に片桐・Juangの提案したMCE学習は特筆に値する），連続音声認識を組み合わせ最適化問題として定式化し，できるだけ正確に解くアルゴリズムを見出すことが，高精度な認識につながるという事実である（2段DP法，One Pass DP法，混合連続出力分布型HMM，triphone，trigram，A*探索など）（**第2部音声認識の情報理論・統計的アプローチによる定式化参照**，p.69）。例えば，混合分布型HMMの音響モデルの場合，ガウス分布数は1万個以上が使用される。また，語彙数が20,000語の言語モデルだとbigramの場合4億通りの，trigramの場合8兆通りのパラメータ（確率値）が必要であるが（確率0を除いても数百万以上のパラメータ），これらの各々に対して10%の値の誤差があっても認識性能に差が出てくる。一方，人間はこれらの値を正確に記憶しているとは到底考えられない。このように工学的アプローチと人間の知覚過程には，量的・質的な差がある。人間の聴覚・知覚過程のモデル化の音声認識への貢献は残念ながらあまりない。

言語モデルについては，現在の主流であるtrigram言語モデルとそのタスク（トピック）による適応化によりほぼ限界に達している。一方，音韻認識率は人間の能力にまだ遠く及ばない。特に雑音環境下やマイクロホンなどの通信回線の違いによる機械の能力は人間と比べて極端に悪い。会話文のような自然な発話（spontaneous speech）に対しては音声認識率は極端に劣化する。これは，自然な発話の言語モデルの構築が難しいこともさながら，発声速度の変動が大きく，発声もあいまいになることが主因である。まだまだ解決すべき研究課題は多い。

なお，大量の音声データベース・言語データベースの共有化以外に，CMUの言語モデルツールキット，HTKのHMMツールキット，IPAプロジェクトの連続音声認識ソフトウェアなどの公開・共用が音声認識研究の進展に寄与してきた事実は見逃せない。

4.3　今後の展開と課題

音声認識への応用で有望な分野として，①カーナビのインタフェース，②モバイルコンピュータのインタフェース，③マルチメディア統合処理（音声検索・音声要約など），④語学教育，などが挙げられる。不完全な音声認識技術でも十分有効だと考えられる③と④の研究は，これからの発展が十分期待できる。

今後の研究方向としては，①人間の聴覚・知覚過程の解明とモデル化による人間に学んだ認識手法の開発，②音声のダイナミクスを表現できる動的システムのモデル化，③自然発話のための発話速度や文脈（品詞列など）に依存した音響モデルや発音辞書のモデル化，④文脈やユーザの意図まで考慮した言語理解，⑤従来の自然言語理解の研究とは一味異なった話し言葉特有のオンライン・リアルタイム性，断片的・漸次的詳細化現象を扱うモデル，⑥異種モーダルとの統合アーキテクチャ，⑦移植性に優れた音声対話システム，⑧ロバストネスの向上などが挙げられる。いずれにしても人間と機械とのコミュニケーション手段としての音声の役割・意義・応用の追求が増々重要となってこよう。

（中川聖一）

5. 音声合成

5.1　はじめに

音声合成の技術は口の機械的モデル作成に端を発し，電気音響理論に基づく音声生成機構の等価電気回路モデル，効率的な音声ディジタル伝送，テキストからの音声合成，マルチメディアシステムの情報出力といった半世紀に渡る展開を遂げてきた。このような音声合成技術の変遷は，動力機械から電気回路，ディジタル電子計算機といった工学全体にわたる技術の歴史の一部であると共に，口の代わりをする器械の完成を目指した，人が持つ音声言語機能の実現の歴史に外ならない。ここでは，音声合成のうち，言語情報を入力として音声波形を出力する技術について，音声言語情報処理機能を電気回路や計算機上に実現するための努力を振り返り，残されている課題を述べる。なお，紙面が限られているので参

照は人名とその当時の所属，関連論文等の発表年を示した。紹介内容が日本，近年の出来事，筆者の関心事に近いものほど細かくなることをご容赦願いたい。

5.2 音声合成器

電気的に音声を生成する方法として，1950年代にMITのStevens，KTHのFantらにより口，声道といった音声波形生成器官が持つ音響的特性を等価回路として再現する声道アナログ型の合成器と，音声出力最終端でのスペクトル共振特性を実現するターミナルアナログ型の合成器が研究され，音声合成器の研究が活発になった[1]。声道アナログ型合成器は，音声生成機構のモデル（AT＆T Bell研究所のCokerら（'67），早大の白井・誉田ら（'76））として開発が引き続き進められたが，次第に音声生成機構のモデル化そのものに関心が移っていった（**第2部X線撮影と音声研究参照**，p. 62）。モデル化や制御の容易さ等から，自由な内容の音声合成を行う目的には，KTHで開発されたOve（'62），MITで開発されたMITalk（'79）[2]等に代表されるように，共振回路を直列に結合したターミナルアナログ型のフォルマント合成器が広く用いられてきた[3]。

音声情報処理への電子計算機の利用が始まると共に，音声の効率的伝送方式として電電公社通研で提案された線形予測分析（LPC）合成系（板倉らのPARCOR合成（'69），LSP合成（'79））は，音響管モデルとして裏付けされた優れた数理的モデルであり，有用な音声スペクトルパラメータ表現を提供した。LPC合成は電子計算機やディジタル信号処理の発達ともあいまって，フォルマント合成に代わるものとして広く使用されてきている（**第2部線形予測分析LPCの発明→PARCOR→LSP参照**，p. 66）。子供用玩具として米国TI社で開発された携帯型英語学習器（'78）は，音声合成部にこの技術を用い，画期的なシステムLSI設計と共に新たな時代の到来を印象付けた（**第2部Speak & SpellとDECtalk参照**，p. 72）。

合成音声の使用が具体化するのに伴い，フォルマント合成やLPC合成に代表される，励振音源特性と声道の共鳴特性を完全に分離したボコーダ型合成器の音声品質向上が課題となった。合成音声の品質向上策としては，励振音源信号のモデル化（Fantら（'85）），声道特性を表すスペクトル概形抽出法の工夫（電総研中島のPSE法（'87）），スペクトル調波構造の表現の工夫（LIMSIのLienardら（'88），IRCAMのRodet（'87），ENSTのStylianouによるHMN法（'96）），波形の位相情報を考慮した分析合成系（NHK清山・都木ら（'92），和歌山大河原のSTRAIGHT法（'96））等が提案されている。また，零位相化したADPCM波形の重ねあわせ（沖の谷頭ら（'85））にヒントを得た，基本周期に同期して一周期波形を重ねあわせて加えるCNETのMoulinesらによるPSOLA合成（'89）は手軽さも受けて広く用いられている（**第2部コーパスベース音声合成参照**，p. 71）。現在ではこのように，高い合成品質を要求される用途に合わせ，合成モデルを用いず，原波形の詳細な情報あるいは原波形そのものを用いる方法も採用されている。

5.3 スペクトル特徴の制御

自由な内容の音声合成では，口のハードウェアモデルとしての音声合成器に対して，機械の口を動かすための音声合成制御ソフトウェアとして，人が音声言語習得に際して獲得した知識を必要とする。「規則による音声合成」と呼ばれるように，実際の音声言語データ分析に基づき，これらの制御知識の規則化とデータ表現が進められた。音色を担うスペクトル特徴の制御の規則化は，Stevens（'56）やOhman（'65）等に代表されるフォルマント遷移モデルに次いで，JSRUのHolmes（'64），AT＆TのRabiner（'68），東北大の粕谷・城戸（'67），東大の藤崎（'70）ら多くの先駆者達によるフォルマント変化のモデル化が行われた。KTHが励振音源信号の精密なモデルを用いて'90頃に例示的に合成した高品質な女声音声は，フォルマントによるスペクトル制御の可能性を示した。

音声合成の商用化のニーズが顕在化するにつれて，MITalk（'79）を皮切りにKlattalk（'81），DECTalk（'83）と自分自身の声を題材に高品質化を徹底的にすすめたKlattのシステム改良（**第2部Speak & SpellとDECtalk参照**，p. 72）とは対照的に，制御データ値の設定を含んだ複雑な手作りの汎用的な規則化は次第に困難になっていった。特に，分析に使用するデータの規定，定量化，評価法が明確でないために，他のエキスパートとの知識共有が難しくなった。この打開策として，明快な基準によりスペクトルを規定する方法がNTT（中嶌らのCOC法（'87）），ATR（筆者らのν-Talk（'88））によって提案され，コーパスベース音声合成技術として進展した[4)~9)]。これらでは，フォルマント周波数に代わって抽出が完全自動なLPC等の数理的モデルによるパラメータを用い，数理的な手法により明確な基準に従い，規定された音声データベースからスペクトルが生成・選択される。音声サンプル選択の基準，合成に必要な音声データの設計法として最終品

質を考慮した選択法（東芝赤嶺（'98）），設計法（KDD河井ら（'99））が考案されている。また，IBM（イタリア）（'89）で試みられたHMMモデルによるスペクトル軌跡の生成も，東工大小林・徳田ら（'96）により大幅な品質向上が図られ，韻律の数理モデルと合わせ制御学習に道が開かれた。また，このような数理的な手法による合成法の利点を生かし，音声認識での話者適応法の利用を用いた声質変換がATRの阿部ら（'88）をはじめとして試みられている。

5.4　韻律特徴の制御

アクセント・イントネーション特性を担う基本周波数の時間制御は，音声学ではトップラインモデルとレンジ変化による説明（例えばAT＆T　Bell研Pierrehumbert，Beckmanらのダウンステップモデル（'88））がなされているが，MIT前田のhat-patternモデル（'76）に見られるように，工学では早くからベースラインモデルが多く用いられてきた。電電公社通研の橋本による点ピッチモデル（'69））を文音声に拡張した箱田らのモデル（'80）や，数多くの文音声，多言語で精力的に実証を続けた東大の藤崎・広瀬らの臨界制動二次系モデル（'71），文基本周波数制御（'84）は構文構造やアクセント属性等の言語情報を入力として動作し，テキストからの音声合成システムに用いられている。更に，これらのモデルを用いた規則の定量化・自動抽出も阪大（山下ら（'91）），ATR（平井・樋口ら（'95））等によって進められている[6]。

一方，音素セグメントの持続時間長（音韻長）の制御はChistovich，Kozhevnikov（'65）によるリズム，テンポといった中枢レベルの抽象的な時間タイミングと調音運動のモデル，Lehiste（'71）による多言語にわたる超分節的な特徴としての共通性の分析，音脚（foot）や拍（mora）といった制御単位，等時性といった制御特性を中心とした長い研究の歴史がある[10]。日本語の音声合成でも，東北大比企（'67）による音韻長の変動分析，電電公社通研佐藤（'77）によるタイミング制御の観点からの制御モデル，東大（樋口・藤崎（'81）），電電公社通研（筆者ら（'84）），ATR（武田・海木ら（'89））の音韻長計算モデルが提案されてきた。この過程で，日本語音声は音響的にはモーラ等長ではなく，高低アクセントの制御とは独立であること等が明確になった。また，韻律を積極的に変える合成用途のために，NHK中村ら（'94）の話速変換法も提案されている。

5.5　今後の展開

機械に人の口の代わりとなる機能を持たせるための音声合成の研究は，口の機能モデルから，合成システム作成のための情報処理モデル，更には人間の持つ音声言語情報の知識表現・学習獲得計算モデルへと発展を遂げつつある。これに従い，音声合成は，音響学，電気工学，情報工学といった技術系の分野から，言語学，音声学，生理学，心理学，認知科学といった幅広い分野にわたった，いわば「人間創生学」とでもいった学問分野の新たな有機的統合を求めている。この一方，原波形が「合成」システムに用いられる現状はフレキシブルな合成器の欠如を如実に物語っており，これまで追求されてきた伝統的な技術項目に対してもより一層の発展が求められている。現在の音声合成が抱えている課題，将来的に展開が期待される課題には次のようなものがあげられる。

（1）　高精度合成器　基本周波数，声質等を自由に制御可能な高品質音声分析合成系

（2）　音声合成システムの最適設計　高品質音声出力システム構築技術として音声コーパス，合成アルゴリズム，主観評価の客観尺度構成，システム設計法としての完成度向上

（3）　自由度の高い合成系の確立　アニメ，擬人化エージェント等への利用を念頭においた人間の持つ声質，話し方のバリエーションの記述，実現

（4）　文生成を含んだ発話生成　言語音声を生成するための情報表現，概念からの音声合成

（5）　合成制御機構のモデル　生成制御機構の解明，生成能力の自動獲得モデル，外国語教育ひいては母国語の理解等への利用を念頭においた音声言語生成認知機構モデル，言葉を覚えるコンピュータ

1) J.L. Flanagan and L.R. Rabiner, Eds., *Speech Synthesis* (Dowden, Hutchingson and Ross, Inc., 1973).
2) J. Allen, M.S. Hunnicutt and D.H. Klatt, *From Text to Speech* (Cambridge Univ. Press, 1987).
3) D.H. Klatt, "Review of text-to-speech synthesis for English," J. Acoust. Soc. Am. **82**, 737-793 (1987).
4) B.S. Atal, J.L. Miller and R.D. Kent, Eds., *Papers in Speech Communication : Speech Processing* (ASA, 1991).
5) G. Bailly and C. Benoit, Eds., *Talking Machines* (Elsevier, 1992).
6) Y. Sagisaka, N. Campbell and N. Higuchi, Eds., *Computing Prosody* (Springer, 1997).
7) R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue, Eds., "Survey of the state of the art in human language technology," Linguist. Comput. **XII・XIII** (1997).
8) J.P.H. van Santen, R.W. Sproat, J.P. Olive and J. Hirshberg, Eds., *Progress in Speech Synthesis* (Springer, 1997).
9) 匂坂芳典，ニック キャンベル，"音声合成のための規則とデータの表現，獲得，評価，" 信学論（掲載予定）.
10) 川崎春子，"音声の時間制御に関するモデルと実測デ

ータ," 音響学会誌 39, 389-397 (1983).

(匂坂芳典)

6. これからの音声研究

6.1 はじめに

近年，音声の分析，符号化，合成，認識などの技術は大きく進展し，多様な応用を生み出してきたが，特に音声認識に関しては，未解決の問題が多数残っている。これらの問題を解決しないと，すぐに実用化の壁にぶつかってしまうであろう。本稿では，これから重点的に研究を進めるべき技術課題，特に21世紀のコミュニケーション環境の中で必要となる技術，周辺の技術的進歩によって可能となる新たな技術などについて述べる。これらの展開に必要な研究体制についても触れる。

6.2 音声認識

音声認識技術の主たる応用は，音声対話とトランスクリプション（ディクテーションを含む）である。音声対話に関しては，音声によるコンピュータシステムとの対話（インタフェース）はいかにあるべきかという点を明確にする必要がある。GUIに代わって，あるいは組み合わせて音声が用いられるためには，単に「音声が人間にとって自然だから」ということではなく，コミュニケーション手段として効果的であることが必須である。人対人のインタフェースに近いことが理想とは限らない。音声による対話が誰にでもできるようにするためには，GUIでアイコンがデファクト標準化されてきたように，音声認識誤りへの対処法を含み，具体的対話方法が標準化される必要がある。そのためには，標準化を意識しながら，実際の応用を目指したシステムを多数構築することが不可欠である。トランスクリプションに関しては，現在の技術では，書き言葉を読み上げた音声なら認識できるが，自由な話し言葉に対しては大幅に性能が低下する。我々の話し言葉の音声に関する知識は，あまりに乏しい。国家的規模のプロジェクトとして，継続的に大量の話し言葉コーパス（データベース）を構築し，多数の研究者が協力して，話し言葉音声の多様な変動に対応できる音素モデルや言語モデルを作り上げることが必須である。トランスクリプションといっても，すべての言葉を文字にするよりも，その内容が分かり利用できる形で出力することが必要である。このためには，意味理解（情報抽出）や要約の技術が必要になる。音声認識の用途を広げるためには，個人差，雑音，歪などへの耐性の向上が重要であることはいうまでもない。音声入力の大きなメリットの一つはハンズフリー入力ができるということであるが，そのためには口から離れたマイクロホンによる性能向上が不可欠である。

コンピュータの小型化，高性能化に伴い，遍在・ウエラブル（ubiquitous/wearable）計算環境の時代を迎える。携帯電話，PDAなどの機能が一体化し，音声認識機能を持った装置を，みんなが携帯あるいは身につけるようになるであろう。それを意識した技術開発をする必要がある。

話者認識への期待も大きいが，セキュリティなどへの実用化のためには，大幅な性能向上が必須である。このためには，多数の話者がそれぞれ時期をおいて複数回発声した音声を大量に集めたデータベースの構築が必要である。

6.3 音声合成

波形接続型の方法により，合成音声の品質は近年著しく向上しているが，個人性の変換（モーフィング）や感情付与に代表される声質変換などへの柔軟性に欠けている。現在の方法で全くできないということではないが，一層の柔軟性を目指した方法の開発が必要である。音声認識の場合と同様に，多数話者による大量の話し言葉音声のデータベースをもとに，何等かの基準に従って，手作業を要することなく自動的に自然な話し言葉音声が合成できる方法の研究が必要である。

6.4 研究体制

これまで，音声生成，知覚などの音声科学は，音声認識，合成などの音声工学の進歩には，ほとんど寄与していない。科学としての音声研究が必要であることは当然であるが，現在の音声工学が抱えている極めて難しい課題を解決するために，音声科学と工学が結びつくことが必要と思われる。最近のIPAの音声プロジェクトの成功に示されるように，特に音声認識に関しては，多数の研究者の協力をベースとする研究体制が不可欠である。多数の研究者によって，大量のデータベースや基本ソフトウェアを共有し，多数のコンピュータの高い能力をフルに活用し，分業し協調する体制が確立できるかどうかが，今後の大きな技術的進歩の鍵を握っている。音声によるインタフェースの研究は，今後マルチメディア・マルチモーダルの枠組みの中で進めていくことが重要である。このためには，コンピュータ工学を含む周辺分野との交流を一層進める必要がある。

(古井貞熙)