1. 音声処理と音響信号のあゆみ

船田哲男†，正会員 春日正男††

キーワード：音声合成・認識，音声伝送・蓄積，マルチチャネルステレオ，ディジタルオーディオ，マルチメディア，ネットワーク音楽

1. まえがき

現在，音声合成・認識がごく普通に生活の場で使われ，また，高品質の音響信号が個人的にも手軽に得られるようになった．ここでは現在に至るまでの音声処理の主要な技術と，特にパッケージメディアを中心とする音響信号の技術の流れについて経緯を概観する．

2. 音声処理のあゆみ

音声は時代を問わず人間どうしのコミュニケーションに手軽に利用されているが，近年は人間と機械のコミュニケーションにも簡便な媒体として利用されるようにもなってきている．ここでは，これらの技術を「音声合成」，「音声伝送・記憶」，「音声認識」の技術に大別して概観する．

2.1 音声合成[1)2)4)5)]

ギリシャあるいはローマ文明の時代から民に話しかける偶像が求められていたように，人の声を発する装置は歴史的に古くから人々の関心が寄せられていた．1879年，Wheatstoneは皮革チューブ管を手で握ることにより種々の声道形態を模擬し，そこへ“ふいご”からリードを通して空気流を送り込んで音声を生成する機械的な装置を作成した．また，1936年にDudleyは音声の生理的な発声機構を等価的な電気回路で置き換えて音声を発生する装置VODERを発表した．これは，有声音は発振器を，無声音は雑音発生器を音源とし，声道機能は10チャネルの帯域フィルタバンクで実現した装置である．数週間の練習により，楽器のようにキーボードとペダル操作で了解可能な音声を発生できるとされるものであった．

音声合成は1970年代に入り実用化されるようになってきたが，初期の頃は，あらかじめ録音された音声波形を必要に応じて適宜つなぎ合わせて再生する録音編集方式であった．これは，波形がPCMなどのディジタル形式で表示されることで実用化が可能になった方式である．しかしこの方式では，ある文脈での音声区間を，違う文脈の音声区間につなぎ合わせても不自然な音声しか再生できないという問題があった．1960年代末から1970年代初めにかけてItakuraとAtalは独立に線形予測分析法を提案した．これは，音声波形の現在値を過去の音声波形の何点かの線形結合で近似する手法である．しかし，ディジタル方式では量子化誤差が避けられないため，線形予測はしばしば発振することが多く，後に安定な合成ができる偏自己相関分析(PARCOR)による分析合成方式が提案された．この方式による音声合成装置として，1970年代後半に英単語発声・つづり方学習器Speak & Spellが市販された．これはアルファベットや登録済みの単語を任意の順で即座に発生でき，テープレコーダではできない機能をもつ機器として注目された．その後，より低ビット化可能でかつ時間軸上での補間特性が良い線スペクトル対(LSP)による合成方式が提案された．

一方，声道アナログ方式やターミナルアナログ(ホルマント形)方式と呼ばれるアナログ方式で音声合成する方法も1970～80年代にかけて提案された．声道アナログ方式は，声道特性を種々の円筒断面をもった音響管をつなぎ合わせて模擬することにより，それと等価な電気回路を音源波(周期波や雑音)で駆動して音声合成する方法である．ホルマント合成方式では，声道の共鳴周波数(ホルマント周波数)と帯域幅をもった複数個の共振回路を接続した電気回路に音源波を入力して音声を合成する．

任意の文章音声を合成する場合には，任意単語の合成や文章音声に付与する韻律(強弱，ピッチや抑揚)情報が必要となる．そのためにも種々の形式の文章音声に対し，有声区間の検出とその区間で時間変化するピッチを抽出するための各種の方法が提案された．これらの方法では，音声波形自体の短時間自己相関関数や，線形予測分析したときの予測残差の自己相関関数，あるいはケプストラム分析したときの高ケフレンシー部でのケプストラム値が利用されている．文を表す文字列からその音声を合成する方式はテキスト音声合成，あるいはいろいろの規則を集積し，それらを適用して合成するという意味で規則合成と呼ばれている．近年は規則によらないで，各音声単位に対する多量のサンプルを用意し，それらの中から必要な音声単位を選択して合成する方式(コーパスベース音声合成)が提案され，文章の文法構造に整合した自然なアクセントや抑揚をもった合成音の生成を目指した研究が進められている．

2.2 音声伝送・記憶[4)5)]

音声伝送は空間的に離れた地点間での音声の送受，音声記憶は時間的に離れた時刻での音声の送受であるという点で共通性をもっている．情報伝送(通信)は古くは“のろし”のように視

† 金沢大学 工学部 情報システム工学科
†† 宇都宮大学 工学部 情報工学科
“History of Speech Processing and Audio Technologies” by Tetsuo Funada (Faculty of Engineering, Kanazawa University, Kanazawa) and Masao Kasuga (Faculty of Engineering, Utsunomiya University, Utsunomiya)

映像情報メディア学会誌 Vol. 54, No. 6, pp. 752～755 (2000)

覚による手段や，“太鼓”のように聴覚による手段でなされている時代があった．しかし，これらの方法では距離に対する制約はもちろんのこと，伝送内容もせいぜい数種類という限られた手段であった．それに対し，音声を直接伝送媒体として利用できることは画期的であり，その出発点が1876年Bellによって発明された電話機といえる．その後，ボコーダ(1939)，ソナグラム(1946)，などの技術が展開され，これらの技術は必然的に，いかに少ない情報量で良い音質をもった音声を伝送できるか，すなわち音声の高能率符号化の技術につながっていく．PCMでは単に音声波形を標本化してディジタル符号で表現するため64～128 kbpsと情報量は大きい．DPCMでは標本間の差分をとることにより音声のダイナミックレンジを下げ，実効的な情報量を減らしている．1970年代中頃に提案されたADPCMでは，量子化レベルを適応的に変更しながら伝送することにより，24～32 kbpsの情報量でも音質の低下を抑えることができる．これらの方式は音声波形を保つような伝送・記憶方式であり，音質も良好であるが，情報量の観点からすると圧縮率はあまり大きくとれない．それに対して，線形予測分析により音源情報と声道情報を分離し，それぞれを個別に伝送・記憶して再合成する分析合成方式で高度な情報圧縮が実現できるようになった．さらに，1970年代後半からベクトル量子化の技法が，偏自己相関方式や線スペクトル対方式のパラメータ表現に利用されるようになり，情報圧縮率は大幅に増大した．これらの方式はADPCMに比較すると音質の面で劣っていたが，1970年代末には音源情報を3種のコードブックで表現したCELPと呼ばれるハイブリッド方式で，音質の向上と低ビットレート化の両方を目指した技術が展開され，4～8 kbpsで良質な伝送・記憶がなされている．近年の携帯電話の急速な普及からして，これらの技術展開は時代の要請といえる．

2.3 音声認識[3)5)6)]

機械による音声の自動認識技術は1950年代に端を発している．1952年Davisにより，特定の話者に対する孤立数字音の認識機械が作成されている．このころの機械は単語音声あるいは単音節中の母音部のスペクトルを調べて認識に利用していた．1959年には話者を特定しない，いわゆる不特定話者に対する母音認識がなされた．1960年，サウンドスペクトルグラフの開発によって，音声スペクトルの時間変化を目で見られるようになり，ある程度の訓練を経れば，その発声内容を視覚で認識することが可能になった．この装置は，聴覚に障害をもった人との音声による通信に適用できる．

1970年代は単音節，あるいは数字音や1単語が認識対象であったが，1980年代に入り，連接した単語あるいは連続音声が認識対象になってくると，高度な技法が応用されるようになってきた．具体的には，標準パターン(テンプレート)との発声時間長差の整合をとるためにダイナミックプログラミングが，認識メカニズムを自動獲得するためにニューラルネットワークが，またスペクトル変動や発声時間変動を吸収するために確率モデルHMMなどが利用されるようになってきた．これらの中でHMMの性能が脚光を浴びるようになり，現在の主流となっている．HMMでは，音声分析して得られる特徴の系列を発生できる確率モデルを推定するため，その尤度を最大にするようにHMMのパラメータを推定することが基本原理となっている．また，音声認識で利用される音声の特徴としては，基本的には周波数軸上の特徴である対数スペクトル(あるいは，その逆フーリエ変換であるケプストラム)が利用される．周波数軸を線形尺度から心理尺度であるメルスケールに変換した特徴や，それらの特徴の時間変化を特徴系列に取り込むことにより認識率の向上が図られている．

さて，言語情報を認識するという目的だけでなく，その音声によって発話者の意図を理解することを目的とした場合があり，これは「音声理解」と呼ばれている．この目的に対しては，音声認識の技術だけでなく，言語モデルや自然言語処理と関連した技術の応用が必要となり，精力的に研究が進められている．一方，話者が特定の話者本人であるかを調べる「話者照合」，あるいは話者があらかじめ登録されている複数話者のうちの誰であるかを識別する「話者認識」を目的とした研究も，1980年代中頃より活発に展開されている．

以上の目的以外にも，雑音が混入している音声や伝送系でひずんだ音声をできるだけもとの音声に復元する「音声強調」，音声から声帯の病変を診断したり，補聴器などの聴覚支援をめざした「医療応用」などが挙げられる．これらの技術は，ハードウェア技術，ディジタル信号処理や確率統計理論の進展に支えられてきたことはいうまでもない．

3. 音響信号のあゆみ

ここでは，音響信号のあゆみとして，まず，レコードを中心に進展してきた技術の流れを概観し，これらの音響信号が最近のマルチメディア時代，さらに，ネットワークを中心とした情報化社会の中で，どのように進展し，利用されてきたかについても述べる．

3.1 蓄音機の登場

音声信号が録音再生できる機械は，1877年，ベルが電話機を実演した年に，エジソンが，スズ箔を直径10 cmばかりの円筒に巻いて音を記録再生する円筒形蓄音機(録音と再生の機能を持った機械：フォノグラフと命名)を初めて発明した．いわゆる「メリーさんの羊」の音声が記録再生された．エジソンは，この他に，電話の実用化に貢献したカーボンマイクロホンを発明し，電灯，映画など，数多くの発明をしている．次いで1885年，ベルがワックス(蜜ろう)で包んだ厚紙(ろう管)を利用して，蓄音機を発明し，発売している．その後1887年，現在のハイファイステレオの原形ともいえる円盤式蓄音器(グラムフォンと命名)をアメリカのバーリナが考案し，新しいレコード製造法であるプレスによる円盤の大量生産法を開発し，これがレコードによる再生音楽の普及の原点となった．ただ，この録音方法は，純機械式の録音方法であり，「ラッパ吹き込み」といわれ，メガホンに向かって直接吹き込むものであった．このため音質はハイファイにはほど遠いものであった．その後，この方法によるレコー

ドは1904年に両面録音方式へと改良された．録音性能としては，800 Hzから2500 Hzの周波数帯域をカバーしていたようである．なお，当時のレコードは，直径10インチ，回転数78 rpmが標準であった．その後，現在の高品質音楽が再生できる方式として，ベル研究所で電気録音機方式が考案され，これによってレコードの音質は飛躍的に改善された．いわゆるオーケストラの音がそれらしく聞こえるようになったわけである．

3.2 モノラルから4チャネルステレオまで[7][8]

電気録音機方式が考案された当時でも，依然として再生機は純機械式のままであった．しかし，1926年になると，真空管，モータで完全に電化された電気式蓄音機が発売された．これには，コーン形ダイナミックスピーカ，マグネティック形ピックアップが採用され，これがいわゆる「SP(Standard Playing)レコード」となって商品化され，現在に至っている．その後，マイクやアンプを利用することにより周波数帯域が100 Hzから9000 Hzまでの音域が録音できるようになった．しかし，これらは，いずれもモノラル方式であり，1つのスピーカ，つまり一点からしか音は再生できなかった．目の前に広がるオーケストラの配置，コンサートホールの雰囲気は再現できなかったわけである．これらのまだ未完成ともいえるモノラル方式の録音再生方式は，次に登場する，いわゆるステレオ，すなわち45/45方式ステレオといわれる33・1/3 rpmのLP(Long Playing)レコード，あるいは45 rpmのEP(Elongated Playing)レコードの開発によって飛躍的に改善されることになった．こうして，音楽信号を2つのチャネルに分離してステレオ再生することにより，方向感や音源の位置感が再生できるようになった．そして，マルチチャネルによる，いわゆる立体再生への道の一歩となったわけである．さてこの方式であるが，これはレコードの溝を45°ずつ傾け，それぞれにLとRとの音を記録する方式である．すなわち，溝をトラッキングする針からの信号を，水平成分と垂直成分とに分け，それぞれ直接音と間接音の成分を記録する．したがって，水平成分が同音量で同質の音が記録されていれば，音は2つのスピーカの中央から聞こえてくる．また，垂直成分は間接音が記録されており，これらの直接音と間接音とが相互に影響し合った合成音となり，音の方向感や，位置感，つまり音像の定位や空間の大きさを表現できることになる．

以上のように，ステレオ方式では，直接音と，複雑に反射してくる間接音の相互作用により，音場感は向上することが実証されてきた．すなわち，チャネル数を増やせばこの効果はさらに向上することが予想され，スピーカの数の分だけ音響信号を記録伝送し，その数の分だけのスピーカ配置によって再生すれば，より立体的に音場が再現できると主張され，4チャネルステレオ方式が開発された．この方式には2つの方式が開発されている．ひとつはディスクリート，つまり独立した4チャネル分の伝送路を有するディスク方式であり，もうひとつは前後左右の音のセパレーション(分離性：チャネルの独立性)を多少犠牲にしながらも4チャネルの信号を伝送できるマトリクスディスク方式とである[7]．これらのマルチチャネルの過程を経た再生音場の制御から，いよいよ信号の品質に注目した方式，つまりディジタルオーディオ技術が登場することになる．

一方，磁気録音技術も，蓄音機技術に20年ほど遅れて，1898年にコペンハーゲンのパウルゼンによって発明されている．最初はピアノ線に磁気的に録音する方式であり，30分の録音再生に成功したものである．この後，磁気テープを記録媒体とした方法が開発され，さらに高周波バイアス法が登場して，今日での磁気録音再生機の標準方式となり，現在に至っている．

一方，1962年には，コンパクトカセットという演奏時間60分の磁気テープを30 gという軽量でカセット化に成功し，今までの業務用を中心とした録音機が民生用として社会生活の中に浸透する契機となっている．

3.3 ディジタルオーディオの開発

ディジタル信号処理技術と，高集積度ディジタルLSIの製造技術の確立とにより，また，信号をディジタル化する際に，サンプリング，量子化，そして高度な誤り訂正ができる符号化等の技術の裏づけにより，安定したディジタル信号が得られるようになった．これを背景に，オーディオ技術にも積極的にPCMディジタル信号処理の利用が検討され，1982年，音楽信号の録音再生を行うディジタルオーディオシステムの代表として，オランダのフィリップス社の提案によるCD(Compact Disc)が開発され，急速に普及したわけである．これによって，今までのSPレコードでは，1 mmの幅に4本，演奏時間は5分程度，次いでLPレコードでは，1 mmの幅に10本，演奏時間は30分程度であったものが，CDでは1 mmの幅に625本，演奏時間は75分程度の信号が記録できるものとなった．一方，この方式では，人間の可聴帯域の上限としての20 kHzを考慮して，ディジタル信号の標本化周波数は44.1 kHzとされ，信号のビット語長は充分なダイナミックレンジと考えられた線形の16ビットが採用された．これによって高品質なディジタル信号が手軽に得られるようになり，CDの民生機器としての飛躍的な普及が始まることになった．また，マルチメディア時代の幕開けともなった．

3.4 パッケージメディアの登場

CDメディアを利用して，音声，映像，テキスト等が混在したCD-I(CD Interactive：対話形CD)が登場し，1986年に規格が提案され，1988年には，音楽，画像，文字，図形，コンピュータソフトおよび数値などを含むマルチメディア情報の記録再生メディアとして規格化された．同時に，この頃コンピュータ産業でもマルチメディアの言葉が，次世代コンピュータのキーワードになるとして紹介され，本格的なマルチメディア時代の幕開けが始まった．さらに1987年には，ハイパーテキストやDVI(Digital Video Interactive)が登場してきている．このように，メディアを一元的に扱うことができるディジタル化が進むことによって，扱いや処理が簡単になった．また，マルチメディア情報の記録再生伝送も容易に実現でき，しかも対話形式でユーザの意図に即応できるなどの特徴も注目され，パッケージメディアと

しての進展が加速されたわけである．

パッケージメディアとして，まずCD-Iがある．これは，オーディオの符号化方式として波形符号化方式のADPCM方式を採用し，対象となる信号の内容に応じてA，B，Cの3つに分類して記録再生している．レベルAは高品質オーディオを，レベルBはFM放送と同等の品質を目指しており，また，レベルCはAM放送と同等の品質である．

次に登場したのが，DAT(Digital Audio Tape)である．これは標本化周波数が48 kHz，44.1 kHz，32 kHzで，信号の語長は16ビットである．アクセスの不便さ，特に著作権の問題などで民生機器としてはあまり普及せず，もっぱらデータの記録用としての使用が多いようである．

次いでディジタルコンパクトカセット(DCC: Digital Compact Casset)，ミニディスク(MD: Mini Disc)などが開発され，現在これらの方式が手軽に高品質オーディオを記録再生できる方法としての主流となっている．

まず，DCCは，従来のアナログコンパクトカセットとの互換性を保ち，ディジタル信号での音楽および文字の録音再生が可能な方式である．DCCは，標本化周波数は，48 kHz，44.1 kHz，32 kHzの3種類があり，オーディオ信号をコンパクトカセットにディジタル記録する符号化方式(PASC: Precision Adaptive Subband Coding)を使用している．この符号化方式は，MPEG-Audioの中で提案されたMUSICAMの帯域分割符号化方式に基づいてDCC用に開発されたものである．入力信号が32個のサブバンド帯域に分割され，各サブバンド信号は，準瞬時圧伸法によって符号化され，一定の伝送レート(192 kbps/チャネル)のデータで約1/4程度に圧縮されて出力される．

次にMDであるが，これはCDの次世代パーソナルオーディオ用としてDCCに対比する形で開発された．最大の特徴は，CDに慣れ親しんできたユーザがクイックランダムアクセスの点で，コンパクトカセットより優れている点を見出している点であろうか．MDの記録方式はEFM変調方式で，光磁気ディスク(MO: Magnet Optical)を採用し，標本化周波数は44.1 kHz，符号化方式は高能率符号化方式としてMPEG-Audioの中で使用しているATRAC(Adaptive Transform Acoustic Coding)を利用して，信号データは約1/5程度に圧縮され，146 kbpsのデータで出力される．なお従来，CDを中心としたディジタルオーディオは1990年代の前半に至り大きく変貌しつつある．これは，現在よりもさらに情報量を増やして，より高品質のオーディオ信号が要求されたことである．つまり，要求される高品質は，ひとつは再生帯域幅であり，もうひとつは1データあたりの量子化語長である．これは，スーパビットマッピング(SBM)に見られるような，マルチビットによる20ビット，あるいは24ビットの量子化語長の採用，96 kHzの標本化周波数の採用など，信号の冗長度を増加させた新たな次世代ディジタルオーディオの動きが現れてきたことである．

3.5 コンピュータと音声音響処理

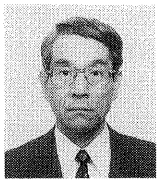
音声信号，音響信号のマルチメディアとしての応用は，電子会議などの音声通信を初め，アニメ，グラフィックス，ゲームなどの音声，3次元音響処理，百科事典，テキストなどの朗読，MIDI(Musical Instrument Digital Interface)，WAVEファイルなど，多種多様な用途があり，ユーザがこれらを組合せて対話的に操作することが特徴となっている．この結果，私達はコンピュータを基本に，何らかの入力装置を使って，表示状態を見ながらシステムと対話して制御していくことが可能となった．例えば，コンピュータ間でグラフィカルな共有環境を実現し，遠隔地の人との共同作業を実現できる方法として，テレコミュニケーションを利用したグループウェアという方法があげられる．これは，計算機をGUI(Graphical User Interface)として利用し，多地点間の通信を目的とした仮想会議や，音楽などの演奏を目的とした再生音場を想定し，仮想的な立体音響環境として頭部伝達関数を実現し，これによるバイノーラル信号を利用して，あらゆる方向への音像定位感の付与により臨場感の向上を図る方式などである[9]．また最近では，音楽をいつでもどこでも演奏会場と同じような雰囲気で聞けるようにと，パッケージメディアに代わってネットワークを利用して音楽配信などを行う，音楽コンテンツの販売という新しいビジネスに展開しようという試みがある[10]．これは音楽信号を1桁以上圧縮できる最近の技術進展の裏づけにより実用化されつつあり，将来楽しみな技術でもある．

以上，音声処理，音響信号のあゆみを概観してきた．将来的には，これらの音声音響信号処理が最近のネットワークを中心とした情報化社会の中で，さらに進展していくことを期待したい．

(2000年3月6日受付)

〔文　献〕

1) J. L. Flanagan: "Voices of Men and Machines", J. Acoust. Soc. Am., **51**, pp. 1375-1387 (1972)
2) B. S. Atal, S. L. Hanauer: "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Am., **50**, 2, pp. 637-655 (1971)
3) 古井："音響・音声工学", 近代科学社 (1992)
4) 広瀬(編)："音声信号処理特集号", J. of Signal Processing, **2**, 6 (1998)
5) T. Chen: "The Past, Present, and Future of Speech Processing", IEEE Signal Processing Magazine, **15**, 3, pp. 24-48 (1998)
6) M. J. Crocker (ed): "Encyclopedia of Acoustics, Vol. 4, John Wiley & Sons, Inc. (1997)
7) "これからのオーディオ", オーディオジャーナル社 (1972)
8) 早坂寿雄："音の歴史", 信学会編, コロナ社 (1991)
9) 北脇信彦編著："音のコミュニケーション工学", コロナ社 (1996)
10) 三宅常之："音楽再生用MP3ソフトが携帯機器のネジ・クギになる", 日経エレクトロニクス, 751, pp. 152-165 (1999)



船田　哲男（ふなだ　てつお）

1966年，金沢大学工学部電子工学科卒業．1971年，名古屋大学大学院博士課程修了．同年，金沢大学工学部講師．現在，同大学教授．生体情報処理，音声情報処理の研究に従事．共著「情報科学の基礎」，「数値解析の基礎」など．

春日　正男（かすが　まさお）

1971年，名古屋大学大学院修士課程修了．同年．日本ビクター(株)入社，総合技術研究所に勤務．1988年，(株)リコー入社，中央研究所に勤務．1995年，宇都宮大学工学部教授，情報工学科に勤務．ディジタル信号処理，感性情報処理および，それらの福祉，医学への応用に関する研究に関心を持つ．工学博士．正会員．