解　説

音声の合成と出力*

市　川　　熹**

1. はじめに

人と人とのコミュニケーションと言うと先ず頭に浮ぶのは音声であろう．人が物事を考えたり思い浮べたりする時に使われる符号系は音声と密接に関係している．言語・思考の中枢と音声の中枢は重なっているとも言われている．マン・マシンシステムの情報媒体として音声が取り上げられるのは極く自然と言わねばなるまい．

音声出力の研究は長い歴史を持つ．情報処理装置の出力として実用化されて以来約十年になる．この間のコンピュータ技術の発達と電話網の拡充に支えられ音声出力を持つシステムは急速に広がろうとしている．特に音声の特徴に注目し真にコスト・パーフォーマンスの良いシステムを構成しようという動きが目立って来ている．

現在の音声出力は基本的にはほとんどの要求に答えられる段階に来ている．問題点は必要な音声データの生産性と任意音声の出力技術である．実用化されているものの大部分は録音編集方式によっている．この方式で出力可能な音声の内容は，あらかじめ録音されている音声の組み合せで実現されるものに限定される．またシステムごとに音声内容が異なり個別に用意しなくてはならない．システムの変更に際しても既に録音されている部分との音質のバランスの取れた追加録音が困難なこと(発声者の体調等)，場合によっては同一人の音声を収録できない事態もありうる．録音編集方式についで実用期に入りつつあるパラメータ編集方式についても事情は同じである．ここに今日音声合成の研究努力が続けられている大きな理由がある．

本解説では，音声出力の特色と現状，応用を中心に現在の音声出力の最大の研究課題である合成の状況も加え紹介する．

* Speech Synthesis and Audio Response Equipment by Akira ICHIKAWA (Hitachi Central Research Laboratory).

** 日立製作所中央研究所第6部

2. 音声出力方式

2.1 方式の分類

発声には，音声波形を形成する手段（喉や口，鼻など）と，それを制御する脳神経指令の双方が必要である．音声出力装置（音声応答装置と呼ぶことが多い）においても，この面から見ることができる．特に制御方式は，任意内容の出力が可能かどいかという視点からも重要な側面である．この視点から制御方式は編集方式と規則方式に大別される．あらかじめ用意されており，意味内容と対応付けられた単語以上の単位を組み合せる制御方式を編集方式，音節や音韻など意味的内容と直接対応していない単位からなる符号系列を入力して規則にしたがい任意の音声を作り出す制御方式を規則方式と呼ぶことにする．

音声波形を形成する手法としては，あらかじめ記録されたなんらかの単位の波形をつなぎ合せる記録再生方式と，波形演算処理により波形を合成して行く合成方式とがある．

音声出力の各種方式は，このように二種の制御方式と二種の波形形成方式の組み合せにより大きく分けられ，さらに制御の処理単位や波形形成のアルゴリズムにより多数の方式に分類することができる．**表-1**（次頁参照）に代表的なものを示す．

しかし，技術的流れから見ると，編集方式は記録再生方式と結びつき録音編集方式として，規則方式は合成方式と結びつき規則合成としてスタートした．その後それぞれの欠点を補うための技術の模索の中から相互にその組み合せを変えた様々な新方式が生れてきた．パラメータ編集や素片方式に含まれる各種の方式がその例である．

合成方式はボコーダ技術（音声の高能率狭帯域伝送技術）の合成部を利用し，分析部からのパラメータの代りに規則により発生されたパラメータやメモリ中に一旦記録されたパラメータを用い音声を合成する．したがって多様なボコーダ技術に対応して様々な合成ア

表-1 音声出力の諸方式

制御方式	編集方式							規則方式						
波形形成方式	記録再成方式				合成方式			記録再成方式				合成方式		
方式名称	音声ファイル	録音編集	素片編集		音声ファイル	パラメータ編集		規則素片				規則合成		
制御単位	文章以上	単語又は句			文章以上	単語又は句		VCV, 音節, 音韻等						
記録単位又は波形形成手法	ファイル	単語又は句	音声素片 自然素片	音声素片 合成素片	音響素片	PARCOR	PARCOR	VOICE EXCITED VOCORDER 等	単音節波形	音声素片 自然素片	音声素片 合成素片	音響素片	PARCOR	スペクトルアナログ	ターミナルアナログ	声道アナログ
波形メモリ又はパラメータメモリ	非常に大	大	中		小	大	中	中	中	小		小	小	小	小	小
制御メモリ	中	小	小		中	小	小									
処理	高速	簡単	簡単		簡単	複雑	複雑	複雑	複雑	複雑		複雑	複雑	複雑	複雑	複雑
語彙数	非常に大	中	大		大	非常に大	大	無限	無限	無限		無限	無限	無限	無限	無限
多重度	小	大	中		中	小	小	小	大	中		中	小	小	小	小
品質	優	優	良		可	優	優	可	可	良		可	良	可	可	可

ルゴリズムを考えることができる．表-1 では音声応答として試みられているものの中の数例を上げたにすぎない．VOICE EXCITED VOCORDER や PARCOR，ターミナルアナログはその例と言って良かろう（声道アナログは逆に音声合成用として発達し最近ボコーダへの適用が検討されはじめた）．パラメータ編集方式はパラメータを記録編集し合成部を制御，音声を出力する方式である．ボコーダについては本小特集の板倉・東倉両氏の「音声分析」を参照されたい．

このように見ると，録音編集技術とボコーダ技術が音声出力技術の大きな柱となっていることがわかる．図-1 に各方式の相互関係を模式的に示す．点線部は直接には音声応答の分野では取り扱わないが，非常に密接な関係にある分野である．音声出力に関する研究開発はこの図の録音編集方式より下の部分を通常対象とし，この範囲の装置を音声応答装置と呼ぶ．

アナウンスマシンは音声応答の範囲外とされるが，利用価値の高い様々な応用がある．また，音声ファイルは静止画情報サービスなどに必須の技術として注目されている．形式的には録音編集の制御単位を極端に大きくしたものである．制御単位は単語から 10 秒以上の話題に拡大され，その種類も数万ファイル以上になる．それ故に技術的には全く異質のものとしてあつかわれる．画像ファイル技術の応用により実現を目ざした検討が多い．

2.2 各種方式の特徴と現状

各方式の得失はすでに試みられている[1]ので，ここでは代表的なものについて簡単に述べることとする．また規則方式の技術的内容については，その概要は章

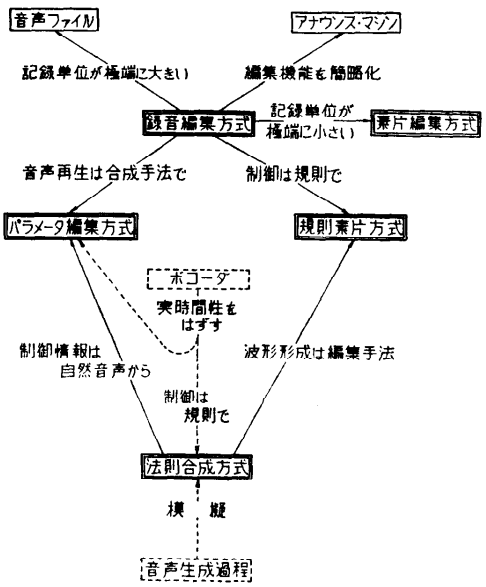


図-1 音声出力方式の相互関係

を改め 4. で述べる．

（1） 録音編集方式

1960 年代にすでに実用化された最も実績のある方式である．多重応答能力が高く，回線当りの経済性をはかることができる．肉声を録音しておき，編集出力するので了解性や自然性に勝れている．記録方式は最近では音質の良いデジタル方式 (PCM) が主流となって来た．また ADPCM を用いた方式は効率良くデジタル信号を圧縮記録した方式として注目される[2]．単語の種類も初期には数十語であったが，現在では 2,000

語以上の大規模なものが実用化されている（たとえば日立 HIVORS H-1300 型，2,048 語 768 回線）．その他，フィルム光学録音を利用した小型のもの（米 Cognitronics 社 630 シリーズ，10～30 語 1 回線），中規模のもの（富士通，256 語 128 回線），フロッピディスクを利用した中規模のもの（日立 H-1100 型，300～1,500 語 16～32 回線）など様々な性能のものが内外のメーカから発売されている．**図-2** にパラメータ編集方式のものと合せ代表的製品の例を示す．

（2） パラメータ編集方式

録音編集方式の記憶容量面からの語彙的限界を拡大することを目的に，ボコーダの情報圧縮能力を利用した方式である．音声は情報圧縮されたパラメータの型で記録されており，出力時に合成器を利用して音声に再合成される．VOICE EXCITED VOCORDER 方式（IBM 7772，生産中止）や PARCOR[3] 方式などがある．PARCOR は電電公社通研の板倉・斉藤両氏により発明された非常に勝れた高能率狭帯域伝送技術である．音質も録音編集方式に比しほとんど遜色がなく，語彙数を飛躍的に拡大させた．すでに実用の段階に入っている（たとえば日立 H-2100 型，約 14,000 語 256 回線）．

（3） 素片編集方式と規則素片方式

素片という呼び方には広義には単語などの大きな単位の波形も含まれるが，最近ではピッチ単位の波形およびそれ以下のものを指す場合が多い．声道の共振情報を持つ音声素片方式[4),5),18)]と，音声素片を単共振波形に分解し一層情報量を圧縮した音響素片方式とがある[6)]．素片方式は高い情報圧縮能力を持つので波形メモリが節約でき，また編集技術による波形形成を行うため多重度も高く取れ経済的である．問題点は女性の声の品質劣化である．

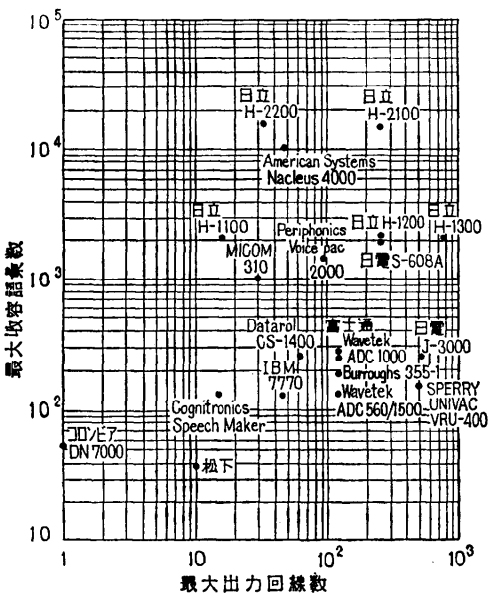


図-2 代表的な音声応答装置

素片編集方式は制御情報を肉声より抽出して用いた限定語彙方式であり，規則素片方式は制御情報を規則によって得る任意音声出力方式である点で基本的に異なる（次の規則合成の一種である）．

（4） 規則合成

規則により制御情報を発生し任意の音声の出力を行う．波形形成手法には様々な方式がある．

規則は言語によっても異なる．現在の所いずれも不完全な状態にある．このため音質も十分ではない．処理量が多いため多重応答向きではない．しかし多重化は行わなくとも半導体技術の発達でメモリやマイコンなどのハードウェアのコストは下がりつつある．むしろ規則のアルゴリズムやソフトウェアの開発コストが問題であろう．

完全に規則化された実用的機種はまだ出現していない．米国 Federal Screw 社の Votrax は半規則方式とも言うべきものである．英語についての簡単な合成規則を，本体とホストコンピュータのソフトに持たせている．あらかじめ利用者が Webster の辞書の発音記号を参考にしながら一定の約束手順にしたがって必要な単語をホストコンピュータに入力する．規則にしたがい音声に変換されたものを聞き取りにより判定修正し，完成した単語の制御情報をホストコンピュータのメモリに記憶させておく．利用時にそれを編集出力する．したがって利用形態は限定語彙のパラメータ編集方式となる．編集方式のデータの生産性改善に焦点をしぼり規則を利用したものと言えよう．この行き方では人手の介入を許すから不完全な規則による音質上の問題の修正や辞書の利用がソフトの許容範囲内では可能であり，利用者も限定した形態が多いので自然性など音質上の要求も甘くて良い．不完全な段階にある規則合成の巧妙な利用法として注目される．

研究段階のものについては **4.** で述べる．

（5） 音声ファイル

電話帯域に制限しても 10 秒の音声はテレビ画面一権程度の情報量となる．かつ数万ものファイルを 1 秒以内にランダムアクセスする能力が必要である．まだ

実用に入っているものはない．現在ビデオディスクの応用が検討されており[7] 近い将来実現されよう．

（6） アナウンス・マシン

ほとんどテープ方式であるが，フロッピィディスクを用いた簡単な編集出力機能を持ったもの（コロンビア DENNON DN 7000 など）もある．

3. 音声出力の応用

3.1 音声出力応用の利点

音声出力装置を持つシステムは他にまねのできない利点を持っている．

出力情報としての音声の特徴には次のようなものが考えられる．

イ．理解が容易かつ伝達速度が速い．

ロ．視線や手足の動作が拘束されない．多数のメータや複雑な図面の一部を詳細に読み取りながら，音声によって同時に他の情報を得たり動作の確認をしたりすることができる．

ハ．人間に対する割り込みの強烈さは他の情報手段ではまねができない．

ニ．複雑な情報や一覧性を要するものをあつかうのには不向きであるが，画像などと組み合せると大きな威力を発揮する．

ホ 人間に密着した媒体であるために，社会制度や習慣・価値観などにより利用形態が左右される面がある．日本における良質な女性の合成音声の実現要求などがその例である．米国では実現が困難で言語情報の伝達に特別のメリットのない女性の声の合成研究はされていない．

ヘ．ハードコピーを残すのには不便である．

次に音声応答装置を利用した代表的システムである電話と情報処理装置を結びつけたもの（利用者からの入力はプッシュホンの押ボタン信号，情報処理装置からの出力は音声応答装置からの音声）を例に主要な効果を上げる．

イ．経済性 電話機を利用できるので端末機器に新たな投資が不要である（通常の電話と共用）．利用者数が多く，各利用者の利用頻度の低いシステムでは特に有利である．

ロ．システムの利用範囲の拡大 電話網の利用によりオンラインシステムの利用範囲を拡大できる．原理的には電話のある所ならばどこからでも居ながらにしてシステムの利用が可能である．そのため新たな応用も生れる．

ハ．省力効果 センタ側での人手（パンチなど）を介さずにデータの自動入力（押ボタン信号による）が可能であり，その先の処理までが自動化可能となる．

ニ．正確さ 押ボタン入力に対する確認をその都度音声出力で行うことにより，素人でも正確に入力でき，システムの信頼性が向上する．

ホ．責任の明確化 データ入力の完全自動化によりデータ入力誤りの責任の所在が明確になる．

ヘ．信頼感の向上 入力の正確さの向上と入力責任の明確化により利用者のサービス提供者に対する信頼感が大幅に向上し，システム運営が円滑になる．

ト．サービスの向上 完全自動化により 24 時間サービスが可能となる．

チ．秘密保持性 暗号コードの利用等により利用者を限定できるとともに，電話機の利用により第三者に見聞きされることなく情報を得ることができる（ディスプレーとの比較）．

問題点としては，入力情報や出力情報が利用者の手元に残らない点であるが，むしろ帳票類をなくすという意味では積極的に評価される点でもある．

3.2 利用形態の特徴

各種の音声出力方式は，見方を変えれば様々な利用形態に答えるべく考案されて来たものと言える．

利用形態には，(A)音声出力のサービスの内容，(B)サービスを受ける人の条件，(C)サービスをする側（設置者）の条件，の3つの側面がある．

利用形態を特徴付ける各項目にはこの3つの側面を持っている．ここでは，最も関係の深い側面ごとにまとめて主要な項目を上げる．

● 先ずサービスの内容に関しては

イ．サービスの種類があらかじめ限定されているか．

ロ．音声出力からの一方的通告（連絡）か．

ハ．サービスの開始がセンタから起動されるか，利用者からの要求によるか．

ニ．データ処理システムの一部か．

ホ．必要語彙数の大小，有限無限．

ヘ．多重出力の要・不要．

● サービスを受ける人の条件には，

ト．限定された範囲か（会員制，企業内等）．

チ．同時不特定多数対象（構内アナウンスなど）か特定対象（電話など）か．

リ．利用頻度の大小．

ヌ．利用時間帯と頻度分布．

● サービス設置者側に関しては，

ル．音声出力に許されるシステム内での負荷量．

ヲ．設置場所．

等が上げられる．

これらの項目内容の組み合せにより出力方式やシステムに要求される能力が決まる．今日ではほとんどの要求になんらかの形で直ちに応じられると言って良い．規則方式や音声ファイルはまだ実用段階ではないが，強いニーズがあれば数年内に一部実用化に入るであろう．

3.3 音声出力の応用例

音声出力を利用したシステムの例を紹介する．国内で実用化されたものは比較的大規模なものが多い．国外では米国が中心である．欧州では西独・チェコなどで導入調査が始まったところである．米国では日本に比し小規模なシステムが多いが様々な分野で音声出力のメリットを生かしている．

（1） 電話座席予約システム[8]

昭和 50 年新幹線博多開業に合せ国鉄がサービスを開始した．音声応答の指示にしたがいプッシュホンより乗車駅等のコード（時刻表に記載）を入力することにより，だれでも居ながらにして座席予約ができる．利用者を限定しない大規模システムの典型的な例である．

（2） 電話勝馬投票システム

日本中央競馬会が，昭和51年10月より運用している．あらかじめ登録された会員が自分の銀行口座の預金残高内の一定限度内で家庭からのプッシュホンにより勝馬を投票する．投票代金および配当金は口座より出し入れされる．会員制のシステムおよび金銭の受授を伴なうシステムの例である．金銭上のトラブルは全く発生しておらず，音声出力を利用したシステムの信頼性の高さを運用実績で証明している．

（3） 銀行システム

米国における音声出力システムの7割程度が銀行業務に用いられている．オンラインシステムの代りに用いられているケースが多い．日本では社内連絡用からスタートし，振込案内や残高照会など顧客へのサービスへと適用業務範囲は急速に拡大するであろう．

（4） オーダ・エントリ

チェーンストアの支店から本部への商品配送依頼などに用いられる．オーダから配送手配まで自動化できることや，24時間サービスができる，人手を中間に介さないためミスが大幅にへる，等々電話と結びつけたシステムとしての利点が最大限発揮できる．米国でも最近最も伸びている分野である．

（5） カタログ販売

24時間受付が可能である．米国ではあまりの便利さに注文が殺倒し商品手配がまに合わず数日で閉鎖した例もある．

（6） 電話問い合せ

不動産・中古車等の在庫・物品の問い合せ．トラック・レンタカー等の配車問い合せなど．

（7） 工程管理等

マシンショップから押しボタン入力で情報を入力することによりパンチカードを追放している．入力データや指示の確認，管理部門からの進行状況問い合せは音声出力による．米大手航空機部品メーカ等で大きな効果を上げている．貨物輸送の運行状況の問い合せ等類似の応用業務も多い．

（8） そ の 他

音声の割り込み機能を生かしたホテルのモーニングコール，オペレータへの異常事態発生の警報，機械装置の操作指示[9] 等多数の応用がある．電話によるダイレクト・メールと呼ばれるものも生れているが，情報公害とならぬ配慮が必要である．

4. 音声の合成

本章では音声出力最大の研究課題である規則合成方式の概要を述べる．詳細は参考文献等を参照されたい[10]．

4.1 音声合成の原理

（1） 声道アナログ方式[11]

音声は声帯からの概周期性振動または口のセバメで生じる雑音を音源とした，口および鼻からなる音響的共鳴系（声道と呼ばれる）の共鳴出力波である．声道アナログ方式では声道の共鳴系を電気的に模擬し音声を合成する．

声道は連続的に面積の変る音響管である．これを図-3(次頁参照)に示すように n 等分し，長さ l の直円筒音響管の縦続接続したもので近似する．

音響管内の音波の伝播は次式で与えられる．

$$\left.\begin{aligned} -\frac{\partial U}{\partial x} &= \frac{A}{\rho c_0^2}\frac{\partial P}{\partial t} \\ -\frac{\partial P}{\partial x} &= \frac{\rho}{A}\frac{\partial U}{\partial t} \end{aligned}\right\} \quad (4.1)$$

ここに　U：体積速度　　P：音圧

ρ：空気密度　　c_0：音速

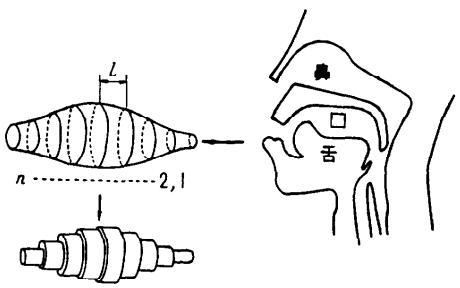


図-3 声道の直円筒音響管縦続接続近似

A：音響管断面積

直円筒管近似に対応して式を (4.1) 差分近似すると

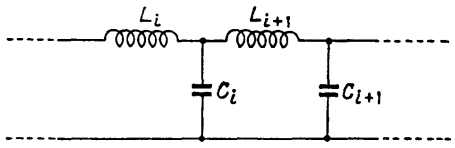
$$\left.\begin{aligned} U_i-U_{i+1}&=l\frac{A_i}{\rho c_0^2}\dot{P}_i=C_i\dot{P}_i \\ P_i-P_{i+1}&=l\frac{\rho}{A_{i+1}}\dot{U}_{i+1}=L_{i+1}\dot{U}_{i+1} \end{aligned}\right\} \quad (4.2)^{*)}$$

したがって音響管は図-4 のように電気回路およびアナコンにより模擬することができる．このような単位回路を声門から唇まで縦続接続し声道を近似する．模擬の方法はこの他様々な変形がある．

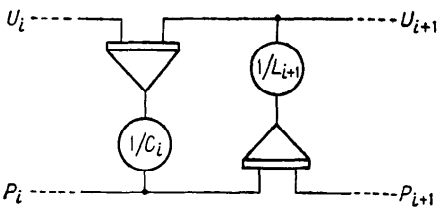
デジタルシミュレーションでは，各直円筒音響管の間の音響インピーダンスの不整合による反射係数を用いたスキャッタ法[2]によることが多い．音響インピーダンス Z_i は

$$Z_i=\frac{\rho c_0}{A_i} \quad (4.3)$$

と近似されるので，反射係数 k_i は



(a) 電気回路



(b) アナコン表示

図-4 声道の模擬回路の例

*) ・は時間微分

**) *は共役な複素数をあらわす．

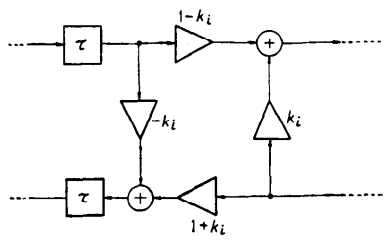


図-5 声道アナログ方式のスキャッタ法によるシミュレーション

$$k_i=\frac{Z_{i+1}-Z_i}{Z_{i+1}+Z_i}=\frac{A_i-A_{i+1}}{A_i+A_{i+1}} \quad (4.4)$$

となる．直円筒での波の伝播遅れ $\tau=l/c_0$ を考慮すると図-5 のようなフローで表現される．

音源や唇の放射特性，鼻腔との接続，声道壁等での損失[13]などのあつかいについてはここでは省略する．

声道レベルの研究は，最近では声帯の発振メカニズム等まで模擬した方式の研究[14]や声道形状のモデル化の研究[15]など，音声の生成過程を解明することを目的とした研究が中心となっている[30]．

(2) ターミナルアナログ方式[16]

音声波形のラプラス変換は

$$P(s)=\frac{k}{s}\left[\frac{1}{1-e^{-sD}}\right]\prod_i^n\frac{s_is_i^*}{(s-s_i)(s-s_i^*)} \quad (4.5)^{**)}$$

ここで $1/(1-e^{-sD})$ は周期 D の音源のくりかえし波形であることを示し，最後の項が声道の共鳴特性を示す．ターミナルアナログ方式は基本的には，この $s_i\cdot s_i^*/(s-s_i)(s-s_i^*)$ なる特性を持つフィルタを n 個(通常は第1～第3ホルマントの3個と高次の補正項) 縦続接続し，音源波形を入力して合成する方式である．z 変換することによってデジタル型の合成も実現される．

(3) スペクトル・アナログ方式

音声のスペクトルエンベロープ特性を適当な周波数間隔で近似する方式である．

(4) 素片方式

音声，特に有声音の部分の波形は相互に良く似た波形が繰り返えしている．図-6 にその例を示す．繰り返えし周期がアクセントやイントネーション，歌のメロディ等を表現するピッチ周期であり，この間の単位波形の持つスペクトル特性が音韻情報を持っている．このような構造に注目し，この単位波形（ピッチ単位

図-6 音声波形の例（女性の声「nana」の「a」の一部）

の音声素片）を肉声から切り出し，素片の尾部を切りつめることによりピッチを変えた音声を作る方式が考えられた(IBM)[17]．長いピッチの制御も可能にするために尾部の後に零をうめる(KDD 研)[5]，線形予測伸長する（日立中研)[4]，零位相化する(日電中研)[35]などの工夫がされた．このような肉声から切り出した波形を利用する方式は，切り出し位置の決定がむつかしく，合成された音声もザラザラした音質となる．これに対し，音声の PARCOR 合成フィルタを求め，そのインパルス応答波形を素片として用いる方式（日立中研)[4] は切り出し位置を絶対的に決める必要がなく，音質も格段に改善される．声道アナログ合成器のインパルス応答波形を利用する方式もある(電総研)[18]．素片の数は類似のものをまとめることにより現実のものの数分の一以下にすることができる[4]．

音響素片方式はターミナルアナログの波形版であり，式(4.5)の個々のフィルタ $s_i \cdot s_i^*/(s-s_1)(s-s_i^*)$ のインパルス応答波形である減衰正弦波を素片としている（日立中研)[6]．この方式では約 60 種の素片で任意の音声を合成することができる．

（5） PARCOR 方式

音韻制御単位 (4.2 参照) ごとに PARCOR 係数を求めておき，規則によって単位間を接続し PARCOR 合成器により音声を合成する(電電公社通研)[19]．前処理として低次の適応逆フィルタによりスペクトル概形を平坦化した音声を PARCOR 分析すると，得られる PARCOR 係数はスキャッタ型声道アナログ方式の反射係数を近似することから[20]，結果的に一種の声道アナログ方式となっているものと考えられる．

PARCOR を変形した合成手法も提案されている(静岡大)[21]．

PARCOR 係数と同じ線形予測係数の一種である自己回帰係数 α は合成フィルタの安定条件がきびしく，パラメータを独立にあつかえないので，規則合成には向かない．

4.2 制御情報（規則）

（1） 規則の構造

規則合成では，音声には時間的に分離量子化される単位（音韻など）の存在を仮定するのが普通である．この単位は連続音声中で具現するとき様々に変形される．規則はこの変形の規則を記述したものであり，出力したい言語情報を入力すると音声が出力するように音声合成装置の制御信号を作る．

連続音声中の音韻の構造を決定する諸要因をモデル的に書くと図-7 のようになる．音声によって伝えられる情報には意図的情報と非意図的情報がある．規則合成では意図的情報のうち，主に言語的情報を伝えることを目的としている．しかし，その他の情報も適切な条件に設定されていないと自然で聞きやすい合成音声を得ることはできない．

音韻性情報は音声の音韻的内容を担うものであり，表現性情報はアクセントやイントネーション，テンポなどにあらわれる文型や強調，喜怒哀楽などの情報を担うものである[22]．

規則への入力は音韻記号とアクセントやイントネーション用の補助記号を用いたものが多い．すなわち，

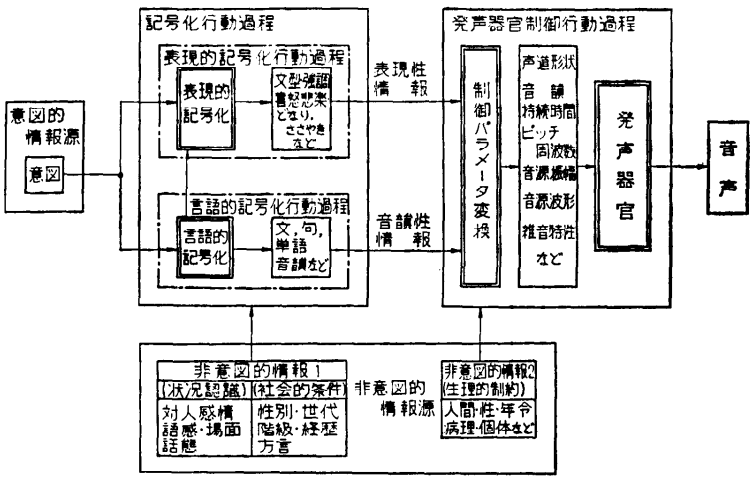


図-7 音声生成過程のモデル

現在の規則の大部分は制御パラメータ変換であり，カナやアルファベット表記入力から音韻記号への対応付け(音韻的記号化)やアクセント・イントネーション・テンポ等の決定規則（表現的記号化）はまだ不備な状況にある．たとえば VOTRAX では制御パラメータ変換を合成器内で処理し，記号化処理をホストコンピュータのソフトに持たせているが，その不備な点を人手の介入修正で補っている．

図-8 に任意単語を合成する規則のマクロな構成例を示す[24]．文章をあつかう場合はこの規則の上に文章の構造を解析決定する部分や，イントネーションなどの決定が加わり，そこから時間構造やピッチ構造に変形を加える．

実用的な側面から見ると，単語，特に固有名詞だけでも任意のものが合成可能であれば編集方式の音声出力と組み合せることにより，ほとんどすべての要求を満すことができる．しかしデータの生産性の問題を解決するためには文章レベルまでの規則化が必要である．

規則の構造や具体的内容は言語によって異なる．一般的規則を作り上げることは当面困難であろう．日本語の規則については電電公社通研などで，英語については Bell 研や英国の JRU で長期に渡って組織的研究が続けられている[19]．

（2） 音韻制御単位

音韻性情報をあつかう単位には様々なレベルのものが考えられる．大きな単位ほど良好な音質が期待されるが，単位の数が大きくなる．手頃な単位は言語によっても異なる．

先ず考えられるのは，音韻や発声の最小単位である音節であろう．音節は日本語では音声学的変種を考慮しても 100 種強で良いが，英語では言語学単位でも 3,500 種以上，音声学的変種（異音）を考慮すると約 10,000 種にもなり実際的ではない．したがって日本語では多少種類が増してもより良い音質の期待できる方向へと検討がされている．たとえば VCV 単位*)(750～1,500 種)がその例である[25]．一方英語では逆に

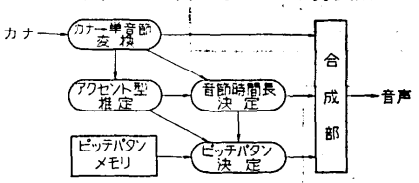


図-8 任意単語合成規則の構成の例

*) V は母音，C は子音．

音節を分解し二音韻の組み合せを基本とした各種の単位 (diad[26], diadic unit[23], diphone[27], demisyllable[28] 等) を用いることにより実用的な数におさえようとしている．

音韻を単位としたものは端末型の合成器など簡便で小型のものを目的とした場合などに用いられる．音韻よりも小さい弁別特徴などを単位としたものは，その音声生成との対応がまだ不明確な点が多く本格的検討はまだと言って良い．

カナなどの入力記号と音韻との対応はかなり規則的である．現実的問題として稼動中の多くの大規模システムではコードでカタカナの大小（ツとッ，ヤとャなど）の区別がないことによるあいまいさが残る．しかし，統計的に処理することにより，あいまいなものの 9 割以上は正しく変換できる．

音韻内の構造の規則については，合成方式により異なるのでここでは省く．

（3） ピ ッ チ

ピッチ情報は日本語ではアクセントや文型（イントネーション），プロミネンス (強調) などの情報を与える．カナやアルファベットの入力から文型やアクセント型を推定することはかなりむつかしい．特に文型では，疑問文等の文型判定はだいたい可能であるが，句間の係り受けの関係によってもイントネーションは異なるので，本格的には言語学的処理が必要になる．

アクセントは日本語の場合，単語や句を一つのまとまったパターンとして知覚させる役割が主であり，意味区別の機能はむしろ例外的である．一つのパターンとしてのまとまり感は聞きやすい音声を作る上で非常に重要である．文字入力からアクセント型を推定することは，会社名とか人名とか固有名詞のカテゴリーを指定すれば 9 割以上は正しくできる[24]．推定誤りの 1 割のうちの大部分も実用上問題のない型が推定される．

アクセントやイントネーションの型から具体的なピッチパターンへの変換には様々なモデルが提案されている．フィルタのステップ入力波形に対する応答波形で近似するもの[29]，点ピッチを指定し直線内挿するもの[31]，などがある．単語のみを合成する場合は種類も有限なので自然音声から抽出したパターンを表引きの型で利用することも可能である[24]．

（4） 時間構造，テンポ

音韻性を与える時間構造の研究[32]から単語[33]や文の時間構造（テンポ）の問題へ研究の中心は移行して来

ている[34]．パターンとしての聞きやすさを与えるパラメータとしてピッチと共に重要である．音声波形の分析的研究から知覚面からの検討へと[34]アプローチの方法も広まり興味ある事実も報告されつつある．

5. 課題と展望

（1） 合成規則

音声出力に対する大半の要求には答えられる段階に来ている．残された大きな課題は，音声データの生産性の問題と任意音声出力技術の開発である．この課題を解決するためには規則合成技術の完成が必要である．

規則合成はすでに一部実用化にふみだし，それなりの効果を上げうる段階に来ている．しかし音声本来の性質である「理解が容易で伝達速度が速い」という点を生かすためには，少なくとも「十分聞きやすい」程度の音質を確保する努力が絶対に必要である．規則合成の研究はその第一段階がほぼ一巡し，学会関係の発表件数も少なくなって来ている．しかし音質改善のためには基本にかえった研究の積み上げが必要である．今日各所で行われている声帯音源メカニズムの研究やテンポの研究，発声生理の研究は規則合成を進める上でも重要な基礎研究である．

（2） 音質評価手法

音声出力装置からの音声の品質評価は，システムの評価や研究開発成果の評価と方向付けのためにも必要なことはいうまでもない．

音声出力装置からの音声はなんらかのレベルで人為的構造が入って来る．録音編集方式においても単語やフレーズのつながり方は人工的であり肉声のそれとは異なる．このような音声に対しては，入力が「人が発声したもの」との仮定のもとに構成されている従来の電話系用の評価手法では十分ではない．音声には了解性と自然性の両面があり，さらに評価者の合成音声に対する「ナレ」の問題や，聞きやすさに効果のある冗長性の要因の相互効果の適正な評価など困難な問題が多い．

（3） 周辺技術

音声出力を十分生かすためには，システムへの効果的かつ経済的な情報入力手段が必要である．押ボタン信号を用いているものが多いが，さらに理想的には音声による入力手段の経済的実現が望まれる．

情報システムが普及するための一つの条件に機密保持機能がある．電話機を端末とするシステムでは，利用者の正当性を判定する情報としては，押ボタン信号による暗号コードの利用か音声波形を利用した個人確認しかない．そのいみで個体からの情報である音声を利用した話者確認技術が求められる．

その他，画像など他の情報媒体との効果的負荷分担の研究や，人間工学的評価も音声出力をより効果的に生かす研究として欠せない．

（4） エレクトロニクス技術の活用

バブルメモリや CCD は音声の記憶媒体としてちょうどマッチした性能を持つ．これら新しい技術の積極的活用が経済性の点からも重要となろう．

（5） 音声出力技術の応用

音声出力は単に効率的情報出力手段というだけでなく，人間性あふれた技術だからこそ大きな効果が期待できる．そこから生れる数多くの技術はハンディキャップ（身障者）用あるいは教育用の機器など生かされるべき分野も多い．米国ではこの面での応用研究もさかんであるが，残念ながら日本では不活発である．この面で努力されている東北大や熊本大，明大には敬意をあらわしたいと思う．

参考文献

1) 中田和男：音声出力装置の現状，計測と制御，Vol. 10, No. 8, p. 579 (昭 46. 8).
石井直樹：電話を用いた情報提供サービス，信学誌，Vol. 60, No. 10, p. 1158 (1977. 10).
2) P. C. Cumminskey et al.: Adaptive Quantization in Differential PCM Coding of Speech, BSTJ, p. 1105, (Sept., 1973).
3) 小池 ほか2：PARCOR 形音声応答装置，研究実用化報告，Vol. 23, No. 10, pp. 2107～2136 (1974).
板倉文忠：新しい音声合成方式，日経エレクトロニクス (1973. 2. 12).
4) 市川 ほか1：音声素片を用いた単音節編集型音声合成方式における音声素片の作成法，信学誌 Vol. 58-D, No. 9, p. 522 (1975. 9).
5) 樽松 ほか1：ピッチ単位音声素片の録音編集による音声合成，昭 45 信学全大 S-3-10.
6) 木村 ほか4：音声応答装置，情報処理，Vol. No. 7, pp. 397～405 (1971).
7) 麻生 哲：画像通信におけるメモリの応用，信学誌，Vol. 60, No. 11, pp. 1337～1341 (1977).
8) 善如寺正雄：プッシュホンによる座席予約システム，情報処理，Vol. 18, No. 10, p. 1050 (1977).
9) "Now, the Talking Computer" Industrial World, pp. 21～24 (July, 1973).
10) 藤村(編)：音声，東京大学出版会 (1972).
11) G. Rosen: A Dynamic Analog Speech Synthe-

sizer, JASA, Vol. 30, pp. 201～209 (1958) ほか.
12) J. L. Jr Kelly et al.: Speech Synthesis, Proc. Speech Comm. Seminar, Stockholm (Sep. 1962).
13) 鈴木誠史：声道のインピーダンスの検討，音響学会音声研究会資料 S-77-26 (1977. 10) ほか.
14) J. L. Flanagan et. al.: Synthesis of Speech from a Dynamic Model at the Vocal Cords and Vocal Tract, BSTJ, Vol. 54, pp. 485-506 (1975) ほか.
15) C. H. Coker: Speech Synthesis with a Perametric Articulatory Model, Speech Symposium, Kyoto, pp. A 4. 1～A 4. 6 (Aug. 1968).
16) G. Fant et al.: Instrumentation for Parametric Synthesis (OVE II), STL-QPSR, Vol. 2, pp. 18～24 (1962).
17) USP 3369077 (IBM).
18) 松井英一：音声素片のピッチ周期編集による多重化音声出力方式：音学講論 1-3-13 (昭 42-11).
19) 佐藤大和：PARCOR-VCV 連鎖を用いた単語音声の合成，音響学会音声研究会資料 (1974-11).
20) 磯道義典：声道モデルと断面積の推定，信学誌 A, Vol. 57-A, pp. 460～466 (1974).
21) 松井英一：CODIC による音声合成方式について，音響学会音声研究会資料 (1976-3).
22) 中山 ほか2：合成音声の自然性に関する基本的考察，音学講論，1-3-7 (昭 42-11).
23) J. P. Olive: Rule Synthesis of Speech form Dyadic Units, 1977 ICASSP Record, p. 568 ほか.
24) 市川 ほか3：音声素片合成，音響学会音声研究会資料 (1975-12).
25) 斉藤 ほか2：音韻連鎖に着目した音声合成システムについて，音学講論，1-3-16 (昭 42-11).
26) G. E. Peterson: Segmentation Tequnigues in Speech Synthesis, JASA. Vol. 30, No. 8, pp. 739～742 (1958).
27) S. E. Esters et al.: Speech Synthesis from Stored Data, IBM Journal, pp. 2～12 (Jan. 1964).
28) O. Fujimura et al.: Demisyllables and Affixes for Speech Synthesis, 9th ICA, I-107, p. 513 (July, 1977).
29) 藤崎 ほか1：日本語単語アクセントの基本周波数パタンとその生成機構のモデル，音学誌. Vol. 27, No. 9, pp. 445～453 (1971).
30) 白井 ほか2：摩擦子音発生過程のモデル化，信学誌 A, Vol. 58-A, No. 6, pp. 345～351 (1976).
31) 橋本 ほか1：ピッチパタンの直線近似の一手法，音響学会音声研究会資料 (1974-7).
32) 比企 ほか：連続音声中の各音韻の持続時間の性質，信学会インホ委資料 (1966-1).
33) 中島 ほか1：単語を構成する音節持続時間の規則化，音学講論，3-2-16 (昭 49-6).
34) 佐藤大和：単語における音韻継続時間と発声のタイミング，音響学会音声研究会資料(1977-10).
35) 伏木田 ほか1：素片編集型音声合成を目的とした自動分析一合成の一方式：音響学会音声研究会資料 (1974-11).

(昭和53年3月2日受付)
(昭和53年4月19日再受付)