

最近の音声合成技術

Recent Trends in Speech Synthesis Technology

音声合成装置のLSI化以来、音声合成技術は実用的に注目されはじめ、エレベーターの案内用、自動車の警報用など広く実用化されつつある。本報告は、音声合成LSI開発の意義とその後の音声合成技術の発展を要約するのが目的である。

まず音声合成と呼ばれているものの内容を分類し、現在主流となっている技術が分析合成法であることを明らかにする。

次いで実用化のレベルではPARCOR方式の音質改善、LSP、ケプストラムボコーダの開発を説明する。また、LSI化についてはCMOS 1チップ化と信号処理汎用LSIの応用に触れる。最後に、真の音声合成となるための規則合成の研究と、音声認識との結びつきによる今後の発展について論じる。

中田和男* Kazuo Nakata

1 緒言

音声合成に必要なものを、人間の音声生成（以下、発話と略す。）過程と対比して示すと、表1に示す三つに要約される。この三つをどのように処理し実現するかによって、音声合成には原理的に種々の方式が考えられる。それぞれに、どのような処理、実現の可能性が考えられるかを表2に示す。

現在大わくの分類として考えられるのは、

(1) 規則合成と分析合成(制御ソフトウェア作成主体の分類)

(2) 生成源符号化と波形符号化(合成ハードウェア主体の分類)

である(表2参照)。生成源符号化というのは、いわゆる音声合成で図1に示すように、音源の生成と調音によるスペクトル形成作用を分離した音声合成装置を使うものである。波形符号化というのは、ADPCM(適応的パルス符号化)のように直接的な波形の情報圧縮符号・復号化を利用するものである。

2 現状の理解

音声合成の現状は、実用製品化のレベルで言えば、録音編集方式から分析合成方式へ移行しつつあり、その大きな推進力は線形予測理論と音声合成装置のLSI化である。別の言い方をすれば、センターの大形周辺機器としての音声応答から、端末機器あるいは小形独立又は付属機器としての音声合成への発展である。

一方、研究開発のレベルで言えば、その主題は分析合成から規則合成へと移行し、また波形符号化の見直しも行なわれ

表1 音声合成に必要な機能 左側に人間での機能を示し、右側にそれに対応した機械側での処理を示す。

人間の発話過程		音 声 合 成	
大 脳	発 話 内 容	入 力	文字系列又は単語番号
神 経	発 話 指 令	制 御	制御情報の作成又は検索(ソフトウェア主体)
口と喉	発 話 運 動	出 力	波 形 生 成(ハードウェア主体)

表2 音声合成方式の分類 音声合成には、任意の音声を出力しようとねらう規則合成方式と、原音声を与えられたときにだけ合成できる分析合成方式がある。

段 階	形 式, 実 現 法	合 成 方 式
入 力	文字系列(文章)	規 則 合 成
制 御	規則による自動作成	
出 力	生成源符号化合成	
入 力	単語, 句の番号	分 析 合 成
制 御	記憶情報*の再生	
出 力	生成源符号化又は波形符号化	

注：* あらかじめ分析によって作成しておく。

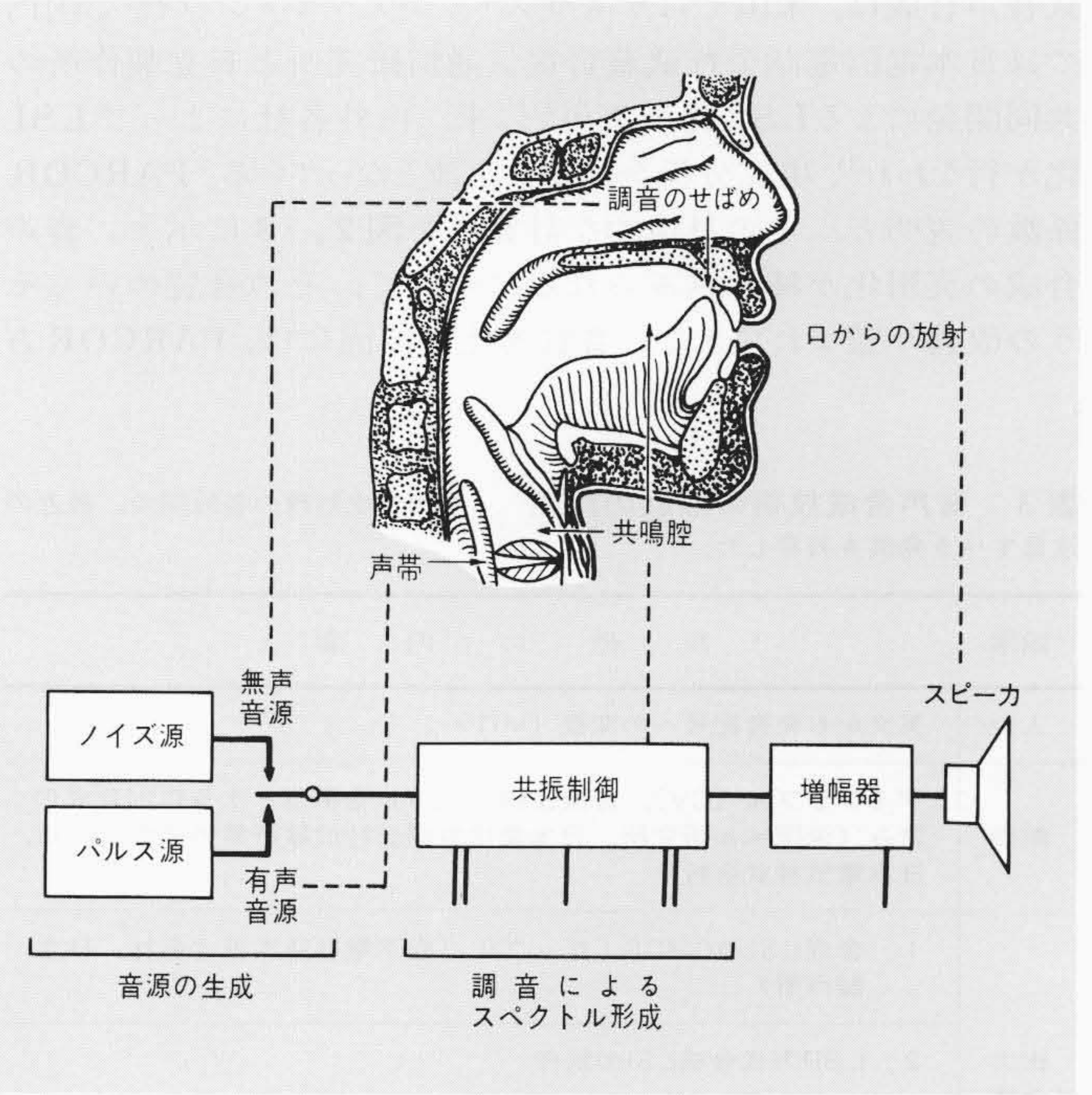


図1 生成源符号化方式の原理 現在の音声合成ハードウェアの基本となっている生成源符号化方式の原理を示す。

* 日立製作所中央研究所 工学博士

ている。

分析合成と波形符号化のいずれを採用するかは、必要メモリのコストと音質との両面から検討して決めるべきである。現状の技術レベルでは、波形符号化による情報圧縮率は多くとも数分の一であり、32~16kビット/秒を必要とする。一方、分析合成によれば必要情報量を2.4~1.2kビット/秒まで下げることができるが、いくらかの音質劣化は避けられない。

したがって、目的とする音声出力に必要な語彙数から、その全語長(秒)を推定し、必要メモリ量を概算し、そのコストが許容できるならば、波形符号化を使ったほうが一般的にいて、ハードウェアが簡単で音質の良い合成音声を得られる。

現在、実用化されている音声合成は、録音編集(波形直接か、波形符号化か)と分析合成に限られており、合成したい語彙について、オリジナルに人間の同一発話内容の音声を必要とする。音声合成という言葉が一般的にもつ「任意の音声」を合成によって作り出せるというニュアンスを満たしていない。

3 最近の技術

音声合成についての最近の注目すべき技術開発の実例を、表1に対応して表3に要約して示す。

音声合成方式について言えば、一つは線形予測分析合成の発展としてのLSP(Line Spectrum Pair: 線スペクトル対)方式の開発と、ケプストラムボコーダ方式の開発である。

また合成装置LSI化の点では、CMOS(相補形MOS)1チップ化による小形、小電力化及び汎用信号処理LSIによる音声合成の試みと波形符号化合成LSIの実現である。一方、研究レベルでは規則合成方式の研究再開が目立つ。一方、PARCOR方式の音質改善の努力も地道に続けられており(当社比較で)、音質改善の実が挙がっている。以下、主要項目について要約する。

3.1 PARCOR方式の音質改善

線形予測分析合成の原理に基づくPARCOR(偏自己相関)方式音声合成は、米国ではテキサス・インスツルメンツ社¹⁾、国内では日本電信電話公社武蔵野電気通信研究所と日立製作所の共同開発によるLSI化の実現²⁾以来、内外各社によってLSI化が行なわれ³⁾、現在分析合成法の主流となっている。PARCOR係数の説明と、その具体的な計算法を図2、3に示す。音声合成の実用化が種々試みられるにつれて、その音質のいっそうの改善が望まれてきた。音質劣化の原因には、PARCOR方

表3 音声合成技術の最近の話題 音声合成処理の各段階で、最近の注目すべき発展を列举した。

段階	・ 技 術 の 内 容
入力	英文から発音記号への変換 (MIT)
制御	デミシラブル (CV)、音韻連鎖 (VCV) を単位とする規則合成の試み (米国ベル研究所, 日本電信電話公社武蔵野電気通信研究所, 日本電気株式会社)
出力 (合成 方式)	1. 合成LSIのCMOS1チップ化 (松下電器産業株式会社, 日立製作所)
	2. LSP方式合成LSIの試作
	3. 信号処理LSIによる合成 (Intel社, 富士通株式会社)
	4. 波形符号化合成LSI (National Semiconductor, 沖電気工業株式会社, 日本電気株式会社)
	5. ケプストラムボコーダの開発 (東京工業大学)

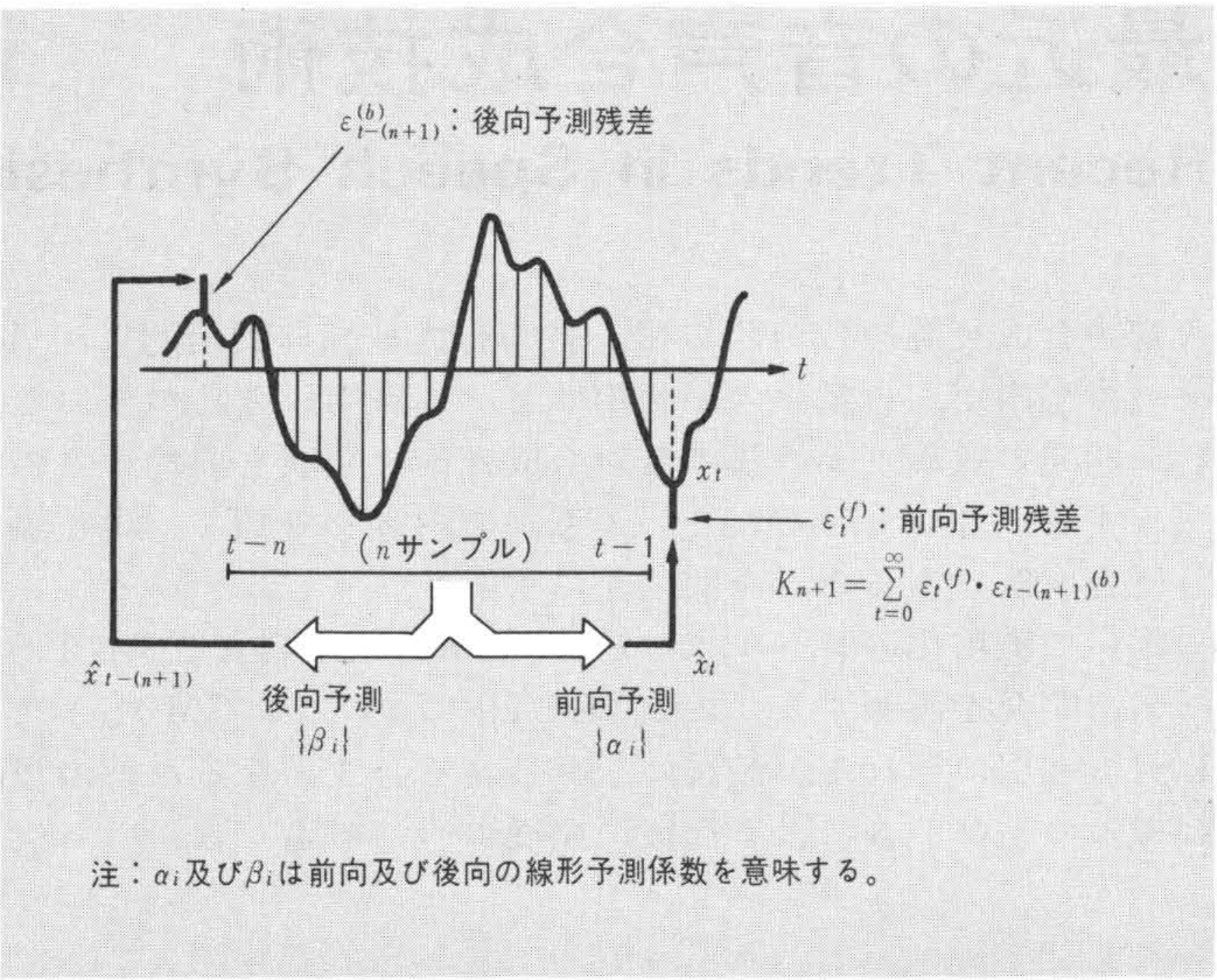


図2 PARCOR(偏自己相関)の説明 PARCOR方式のパラメータであるPARCOR係数の物理的なイメージの説明を示す。

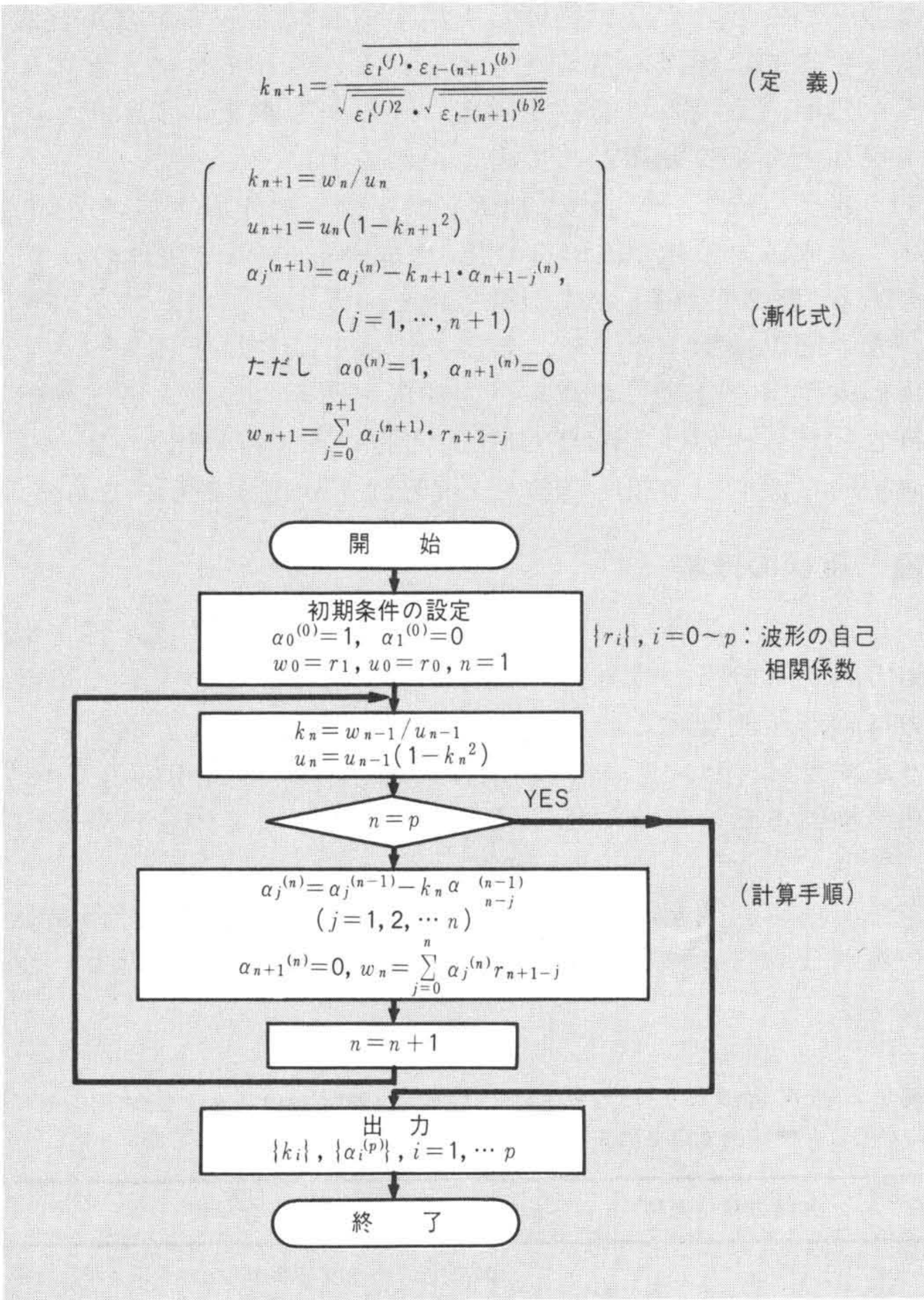


図3 PARCOR係数の計算法 図2に示した原理を、具体的にどのように計算するか標準的な手順を示す。

式に固有なものと、より一般的に、生成源符号化方式に固有なものと考えられる。前者の一つとしてはPARCOR係数のフレーム間線形補間処理(例えば20msごとに更新されるデータ間で2.5ms刻みの線形補間を行なう。)によるスペクトルひずみの増大が指摘され、後者の一因としては、パルスとノ

イズによって音源波形を単純にモデル化し、音源と調音を独立として、その相互作用を無視した近似の粗さが指摘されている。

音質改善は、これらの原因によるスペクトルひずみを軽減する方向に沿って行なわれている。その一例として、日立製作所の試みた改善策の幾つかを以下に要約して説明する。

(1) 線形補間特性の改善

1フレーム20ms区間を2.5ms単位に分割した8点の範囲で、隣接フレームの代表点との間で線形補間を行なったとき、総合的なスペクトルひずみの総和が最小となるような代表点を選ぶことによって軽減することができる。なおこの代表点の選択は、ダイナミックプログラミングの手法を使って比較的簡単に求めることができる。

(2) 音源特性の改善

基本的にはより実体的な音源モデルを開発すべきもの⁴⁾と思われるが、それにはなお原理的な検討、解析と生理的なデータが必要であり、ここではより簡便な対策を述べる。

その一つは積分音源の使用であり、分析に先立つ高域強調(微分)処理と結合させて、ある範囲ではあるが、スペクトルの総合周波数特性を制御することができる。

もう一つは残差音源の使用³⁾である。代表又は平均残差音源波形をいかに作成するかがポイントであり、この点についても日立製作所は独自の手法を開発し、音質改善の効果を確認している。

3.2 新しい音声合成方式

(1) LSP(線スペクトル対)方式⁵⁾

制御情報量が少ないときのPARCOR方式の欠点(音質劣化の原因)として指摘されるのが、PARCOR係数の線形補間によるスペクトルひずみの増大であり、それを低減するため開発されたのがLSP方式である。その分析の手順を図4に、合成回路の構成を図5に示す。PARCOR方式で線形補間特性が悪いのは、補間されるPARCOR係数が、音声波形あるいはそのスペクトルの直接的な物理的特徴量と単純な1対1対応をしていないためである。LSP方式の制御パラメータである線

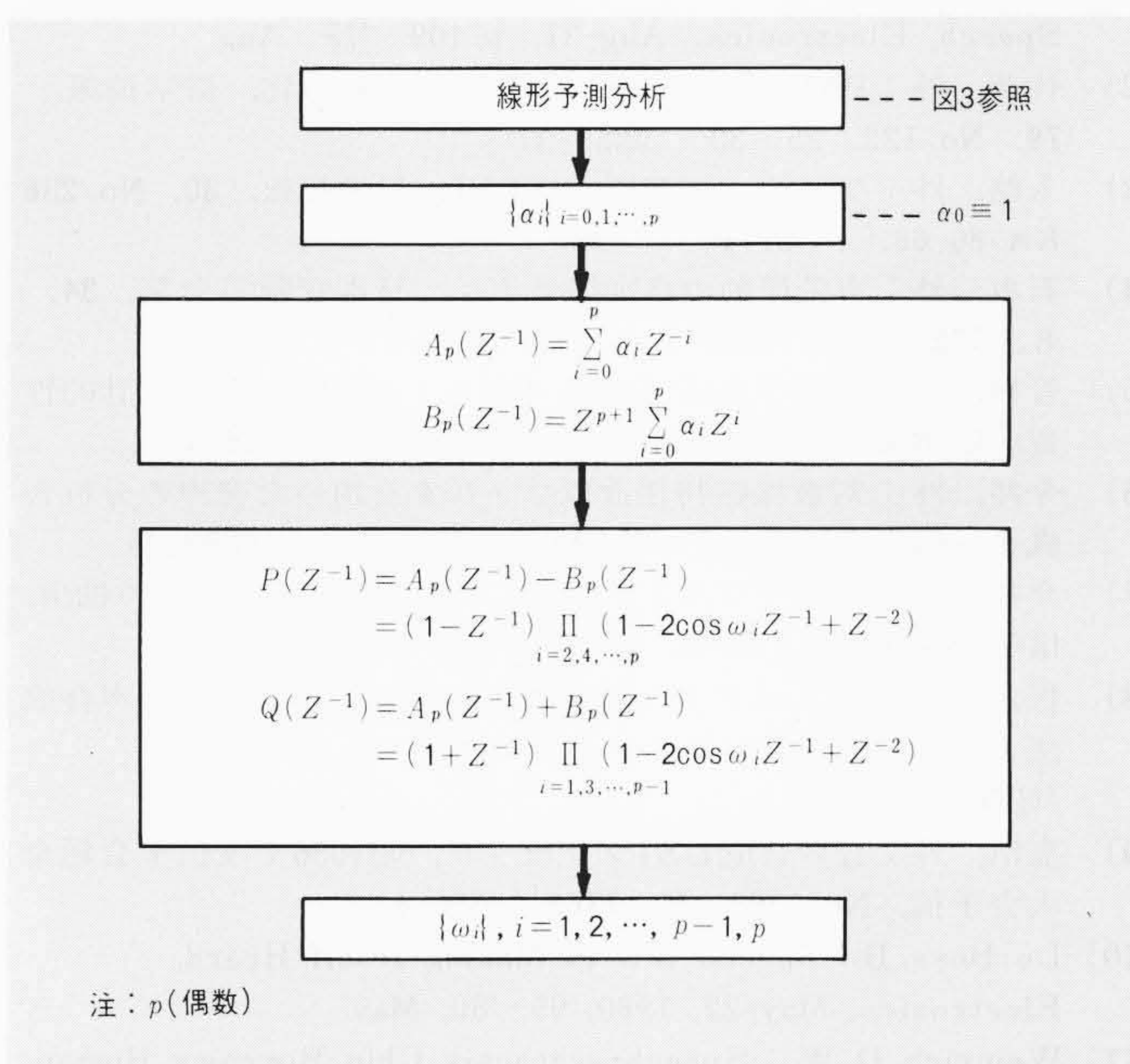


図4 線スペクトルパラメータ $\{\omega_i\}$ の求め方 PARCOR方式の改良を目指して、日本電信電話公社武蔵野電気通信研究所で開発された線スペクトル対の求め方を示す。

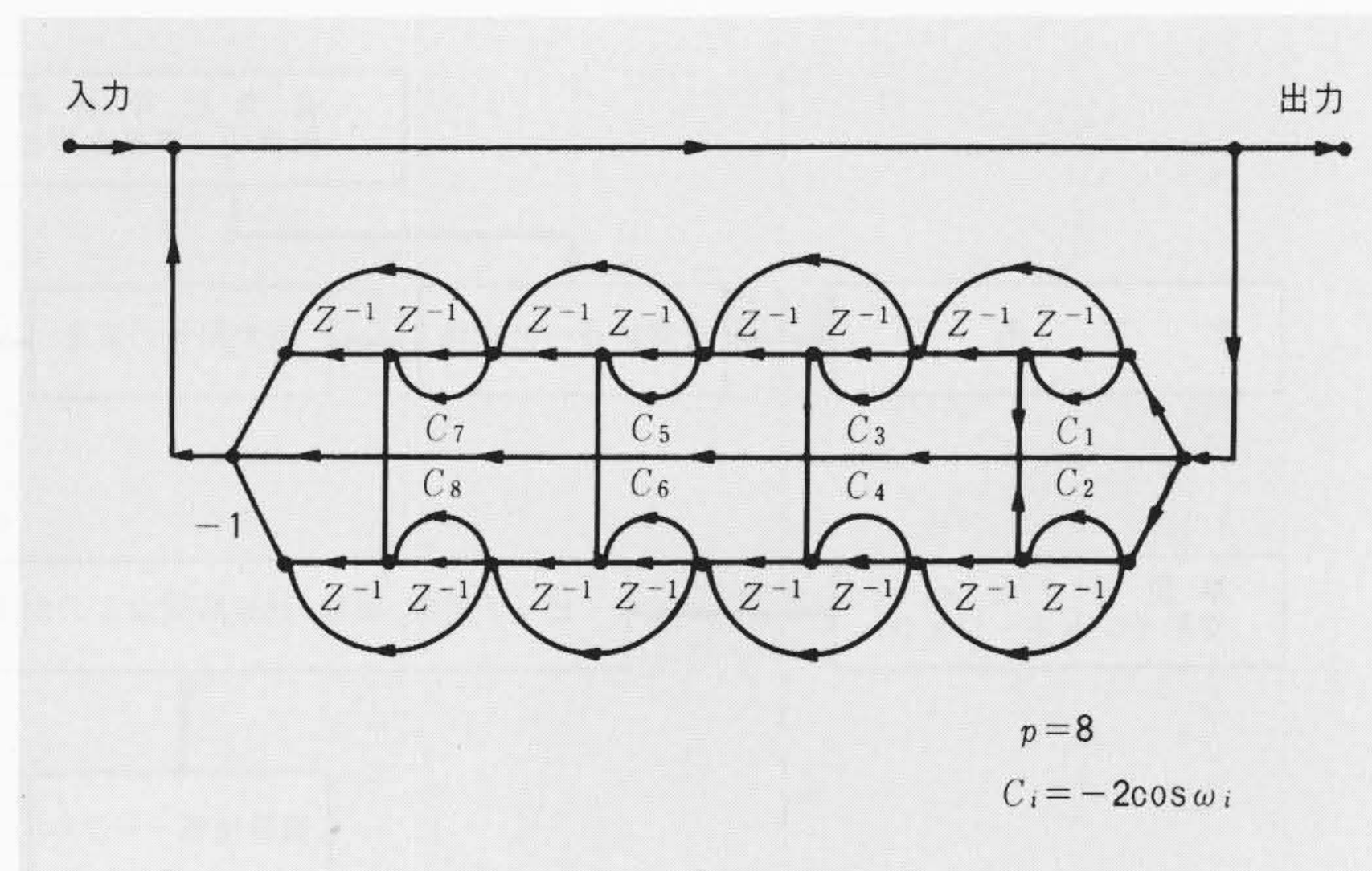


図5 LSP合成デジタルフィルタの構成 線スペクトル対(LSP)パラメータ $\{\omega_i\}$ による音声合成の処理を示したもので、そのまま合成用デジタルフィルタの構成となっている。

スペクトル対係数は、音声スペクトルのホルマント周波数に直接関係づけられる特徴量であり、したがって、線形補間によるスペクトルひずみが少ない。PARCOR方式に対する改善効果は、制御情報量に換算して約85%と報告されている⁵⁾。

(2) ケプストラムボコーダ⁶⁾

PARCOR方式やLSP方式では、分析合成の対象とする音声波形のスペクトルを、全極形モデルで記述できるものと仮定して処理している。本方式では、音声波形の実際の対数スペクトルエンベロープ(デシベルで表現されたスペクトル特性)を、2乗平均誤差最小で最良近似することのできるデジタルフィルタを構成して合成する。

そのフィルタの係数は、合成すべき音声スペクトルのケプストラムの低次成分から決定される。ケプストラムによるスペクトルエンベロープの抽出についても改良を加え⁷⁾、PARCORやLSPと同程度の制御情報量で、同程度の合成音声が出ることが実験的に立証されている⁶⁾。

3.3 合成装置のLSI化

(1) PARCOR合成のCMOS 1チップ化

PARCOR方式は当初PMOS(PチャネルMetal Oxide Semiconductor)を使ってLSI化されたが、小形・小電力化の要求からMOS、特にCMOS(相補形Metal Oxide Semiconductor)による1チップ化が進んでいる。日立製作所の新製品によれば、制御情報用ROM(Read Only Memory)32kビットを内蔵し(語彙の拡大は外部専用ROMの外付けで可能)、1チップで最大63語まで出力できる。消費電力はPMOSの場合に比べ音声出力時で数十分の一に低減されている。

(2) LSP方式のLSI化

日本電信電話公社武蔵野電気通信研究所で開発されたLSP方式合成のLSI化が試みられた⁸⁾。

(3) 汎用信号処理LSIによる音声合成装置の実現

PARCOR方式⁹⁾、ターミナルアナログ形¹⁰⁾、波形符号化方式^{11),12)}など各種が試みられている。

4 音声合成技術の今後の課題

(1) 規則合成の実用化

文字系列入力から、明瞭で自然な任意の合成音声を作り出されて、はじめて本当の音声合成といえる。LSI化による音声合成ハードウェアの実現が当然となった現在、音声合成の研究は再び任意音声合成のための規則の開発を目指しており、内外とも研究が盛んである。その概要を図6に示す。

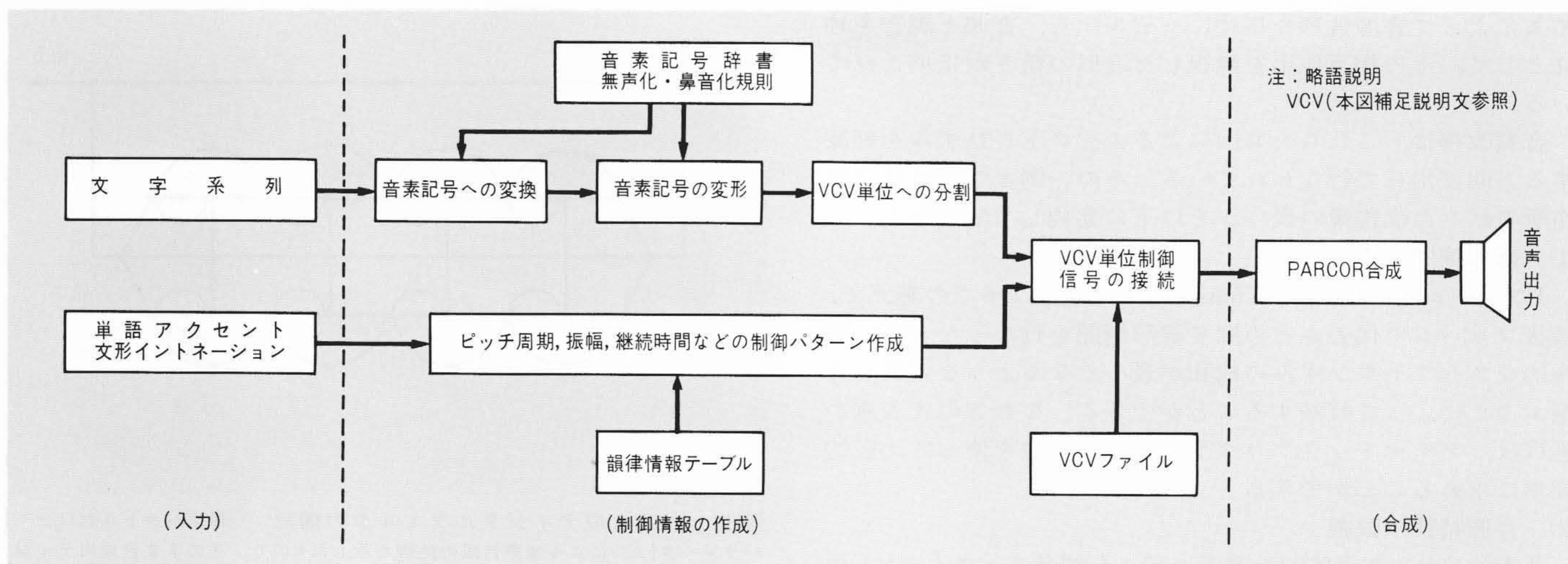


図6 VCV単位PARCOR合成による規則合成システムの構成 文字系列から任意の音声を合成しようとする規則合成の考え方を、VCVを単位とし、PARCOR方式を具体的ハードウェアとしたときの姿で示す。VCVというのは、母音・子音・母音のつながりを考えたもので、音声の一つの構成単位である。

この規則がマイクロコンピュータで実時間処理できるほどに集約化されれば、ワードプロセッサ用の音声読み合わせ機能など音声合成の応用分野は飛躍的に拡大されるものと思われる。

(2) 分析合成のための実時間分析

現用音声合成の主流となりつつある分析合成方式の弱点の一つは、合成のための制御情報を作るための分析が複雑であり、ミニコンピュータ以上の計算能力と、ある程度の経験とノウハウ(各パラメータの符号化フォーマットなど)を必要とすることである。このためユーザーが自由に自分の欲しい音声を作り出すことができず、メーカー側でも小口ユーザーへのサービスコストの負担に耐えられない状況にあり、音声合成チップ応用拡大の一つの障壁となっている。マイクロコンピュータシステム、例えば日立製作所のベーシックマスタにアダプタとして接続できるような実時間音声分析部(できればLSI化された一ボード)が実現され、ユーザーが自分で合成したい原音声を分析し、マイクロコンピュータによる簡単な手直しを経て所要フォーマットに符号化してメモリに書き込めるようになれば、マイクロコンピュータのユーザーを通じて音声合成の応用分野が更に拡大され、音声合成は本当に我々に身近なものとなる。

なおこの技術は、音声波形の低ビットレートによる情報圧縮ディジタル伝送にも直結するものであり、その意味で重要な意味をもつ。

(3) 音声入力(認識)との結合

通信は本来双方向性であってはいじめて十分にその機能を発揮し、有用となる。音声合成は機械から人間への情報伝達ではあるが、人間から機械への情報伝達は音声認識の実現をまわってはいじめて可能となる。したがって、音声認識と結合して双方向性となることによって、音声合成の用途は飛躍的に広がるものと期待される。

5 結 言

音声出力の利点は、常時注意していなくても、必要なときにすばやく注意を喚起することができ、しかもその内容を即時に理解できることにある。また、ラインプリンタ用紙をむだにすることもなく、高価なディスプレイ装置も必要としない。また音声合成LSIの場合、制御情報がICメモリに記憶さ

れているから、その中の任意の一つにランダムアクセスすることが可能である。更に、テープレコーダのように回転や摩擦部分を含まないから、長時間使用や繰返し使用に耐える。

現在までに発表された音声合成LSIの応用製品をみると、上記の利点を活用したものがほとんどであることがうかがわれる。今後ともこうした応用は拡大していくものと考えられる。興味半分の音声ブームがしだいにおさまリ、コストパフォーマンスをしっかりと評価した実用化が、これから本格化するものと期待される。それには、ハードウェアはLSI化されたことでほとんど大幅な改良は望めないもので、音質の向上が今後の決め手であり、合成のための分析のコストが次の課題となる。これらにうまく対処できたメーカーが生き残っていくものと予測される。

参考文献

- 1) Wiggins, R. et al.: Three-chip System Synthesizes Human Speech, Electronics, Aug-31, p.109 '78 Aug.
- 2) 佐藤, 外: PARCOR 形音声合成装置のLSI化, 信学技報, 79, No.122, 25~30 (昭55-3)
- 3) 入路, 外: ワンチップ音声合成LSI, 信学技報, 80, No.236 EA-80-68, (昭81-1)
- 4) 石坂, 外: 声帯振動の自励振モデル, 日本音響学会誌, 34, 8, 122~131 (昭53-3)
- 5) 管村, 外: 線形予測係数の線スペクトル表現とその統計的性質, 信学会論文誌, J 64-A, 4, 323~330 (昭56-4)
- 6) 今井, 外: 対数振幅特性近似フィルタを用いた音声の分析合成系, 信学会論文誌, J 61-A, 6, 527~534 (昭53-6)
- 7) 今井, 外: 改良ケプストラム法によるスペクトル包絡の抽出, 信学会論文誌, J 62-A, 4, 217~223 (昭54-4)
- 8) 板倉, 外: 線スペクトル周波数をパラメータとした音声合成法とそのLSI化, 日経エレクトロニクス, 128~158 (昭56-2)
- 9) 島田, 外: 音声合成LSIプロセッサ, 昭和56年度信学会総全大会予稿, No.1365, No.1366 (昭56-4)
- 10) Le Boss, B: Speech I/O is Making Itself Heard, Electronics, May-22, 1980, 95, '80, May
- 11) Weinrich D. W.: Speech-synthesis Chip Borrows Human Intonation, Electronics, April-10, 1980, 113, '80, April
- 12) 森戸, 外: 波形対称化による音声の帯域圧縮, 音響学会講演論文, 1-7-13, (昭56-5)