音声からの音源情報の抽出とモデル化に関する研究

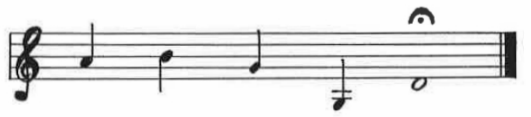
真　野　　淳[1)]

1988 年 3 月 3 日

1)　慶応義塾大学大学院理工学研究科電気工学専攻

目次

緒　　論



―これは、映画「未知との遭遇」において、人間が宇宙からの音声信号を受信した時の音階である。そして、異星人との接近遭遇を果たすべくこの音階を奏で、劇的な接近遭遇を実現したクライマックスシーンを、今でも鮮明に覚えている。もしかしたらこの音階こそ、異星人にとっての音声であったのかもしれない。―

〔１〕　音声と音声信号処理

人間の音声は、通信、すなわち communication のためにある。Shannon[1] の提唱した情報理論の概念に基けば、音声はメッセージの内容または情報で表現することができる。そして、このメッセージ情報を運ぶ音響的な波形が音声信号であり、音声による通信は、この音声信号を介して行われる。この時、メッセージ情報は、大脳から調音機構を制御する一群の神経信号に変換され、調音機構によって音声波形として発声される。このように発声された音声波は、空気中を伝わって人間の耳に到達し、聴覚器官から聴神経を介して聞き手の大脳に伝えられ音声波に含まれる音声情報が認識・理解される。

そして現代の人類は、このような音声による communication の距離的・時間的な拡大、蓄積、あるいは反復を可能にする技術、すなわち音声信号処理という優れた科学技術を発展させてきた。音声信号処理は、1. 音声の情報源を測定、あるいは観測することにより音声信号を取得し、2. その音声信号を分析、表現し、3. 表現された音声信号を変換し、4. 音声情報の抽出と利用を行う過程において、2. 及び 3. の音声信号の表現・変換を行う処理であると定義できる。

ここで言う音声情報とは、大きく分類して２つの情報から成っている。すなわち、離散的な有限の記号の集合である音素 (phoneme) によって表現される大脳レベルでのメッセージ情報と、調音機構を介して発声された連続的な音声波に含まれる肉声性、大きさ、及び発声者の個人性、感情などの自然性に関する情報である。このように、音声情報は非常に多くの多様性を有し、これは音声生成機構の階層性と心理、生理、物理機構の相互作用によるものである。従って、音声信号処理に基く音声通信システムの構築にあたっては、音声の伝送・蓄積・反復を行うために、いかにしてメッセージ情報を保存し、自然性が劣化しないように音声信号を融通性ある形式で表現・変換するかに留意する必要がある。

〔２〕　高度情報化社会における音声情報の役割と音声情報圧縮

音声は人間の communication における重要な手段であることより、社会における音声情報の役割も非常に大きなものがある。特に、OA 化の進んだ現代においては、数字、文字と共に電話での音声通信は必要不可欠な通信手段である。特に近年、ディジタル信号処理技術と LSI 技術に支えられたハードの驚異的な高性能化、小型化に伴い、ディジタル伝走路の拡大及び ISDN(Integrated Services Digital Network) 構想などに基く各種サービスのディジタル統合化が、我が国を初め諸外国で急速に進んでいる。このようなネットワークのディジタル化には、次に示すような優れた利点がある。すなわち、

1. ディジタル信号処理においては、アナログに比較して極めて技巧をこらした信号処理を実現でき、結果的にアナログよりも能率的かつ経済的な伝走路を実現できる。
2. 情報の蓄積・交換・処理に適しており、サービスの高度化及び新サービスへの展開が期待できる。
3. ディジタル信号の符号化を行うことによって非常に雑音の多いチャネルでも正確に伝送することができ、中継性能が高い。
4. ディジタル信号の多重化を行うことで伝送と交換の機能を統合できる。
5. 音声・画像・データを問わずに統一的な形式で伝送でき、通信網を選ばない。
6. 通信内容の秘密保持・誤り対策・保護対策が容易である。

などの利点がある。

ディジタル伝送では、その基本尺度は使用回線の単位時間あたりの情報の伝送量（ビットレート）で決まるため、ネットワークの効率的かつ経済的利用と通信需要に対応するためにも、伝送情報の圧縮技術の早急な開発が必要になってくる。現段階でのネットワークのディジタル化においては、音声信号は１回線あたり 64Kbits/sec を基本として 32Kbits/sec、16Kbits/sec という階層を形成しており、一方、データ通信を中心とした音声以外のディジタル信号は、9.6Kbits/sec を基本として 4.8Kbits/sec、2.4Kbits/sec という階層を形成している。しかし、人間の通信手段としての音声が、速報性・指示性・警報性といった機能を要求され、社会においては多くの場合、事務連絡・活動指示といった通信に使用されることを考えると、音声に対する前記の必要情報伝送量が他の情報に比べて格段に多く、低ビットレート化によるコストダウンが緊急に必要な課題である。

音声信号を 2.4〜9.6Kbits/sec の低・中ビットレートで高能率符号化できれば、移動無線等における音声の中品質伝送等が可能になる。

〔３〕　音声の高能率符号化と本研究の位置づけ

音声信号の低・中ビットレートでの高能率符号化技術は、ここ数年で目ざましい発達を遂げた Digital Signal Processor (DSP) と呼ばれる信号処理用 LSI チップにより、これまで実用上複雑と考えられてきた符号化アルゴリズムも比較的容易にハードウエア化が可能な状況にあり、回線の有効利用や新しい複合サービスの早期実現のために、非常に盛んな研究が行われている。

低・中ビットレートにおける高能率音声符号化方式としては各種のものが提案されているが [2]、大きく分けて２つの方式がある。

第一の方式は、波形符号化方式である。これは量子化誤差の範囲内で音声波形をできる限り忠実に符号化する方式であり、原理的には音声信号でなくても適用することが可能である。波形符号化方式は、16Kbits/sec 以上の高ビットレートにおいては非常に高い音質を有する。しかし、16Kbits/sec 以下の中ビットレートになると音質が明らかに劣化し始め、9.6Kbits/sec 以下の低ビットレートになるとその音質は極端に低下する。

第二の方式は、音声信号を音声生成モデルに基いて分析しパラメータを抽出し符号化を行う分析合成符号化方式である。音声信号の大幅な情報圧縮を行うためには、音声信号からの音声情報の効率の良い抽出が必要であり、そのためには音声生成モデルの導入が不可欠である。生成モデルに基いた代表的かつ実用的な分析合成方式としては線形予測モデルに基く符号化方式があり、LPC(Linear Prediction Coding) 分析合成方式と呼ばれている。現在、有効とされている低・中ビットレート以下の符号化方式は、大部分が LPC 分析合成方式を基本として 音源に相当する残差信号を時間域でモデル化したり、低域成分をベースバンドに変換した後、波形符号化をすることにより符号化効率の向上を図っている。

本研究は、低・中ビットレートでの音声生成モデルに基く分析・合成システムにおける音源情報のモデル化の一方式として位置づけされ、その目的は、LPC 分析合成方式における音源情報の正確な抽出、並びに音源生成モデルに基く音源の新たなモデル化と、それを用いた LPC 分析合成系の高音質・高能率化を達成することにある。

〔４〕　本研究の概要

本研究は、まず第１章において音声生成モデルに基く高能率符号化について概観し、LPC 分析合成方式における音源情報の抽出とモデル化に関する従来の諸研究についてその特徴と問題点を明らかにする。

第２章では、LPC 分析合成方式における音源情報の正確な抽出を目的とした、LPC 分析合成の原理を用いた残差信号の振幅包絡特性からの音源情報抽出法について述べる。

第３章では、LPC 分析合成方式における音源情報の新しいモデル化法として、LPC 有声音残差をピッチ同期分析することによって得た振幅スペクトルが極めて特徴的な零点特性を有することを示し、このスペクトルに対する逆 LPC 分析によるモデル化法について提案し、LPC 分析合成系の高音質化の可能性について論じている。従来、難しいとされていた LPC 有声音残差のモデル化について、ピッチ同期分析を導入することにより周波数域での音源生成モデルを仮定できることを示している。

第４章では、第３章で提案した音源モデルに基いて LPC 有声音残差を高能率に符号化する方式として、ピッチ同期メル逆 LSP 分析合成方式について提案し、LPC 分析合成系と組み合わせることにより低ビットレートで高能率かつ高音質な音声の符号化及び合成を実現できることを示す。

第５章では、各章の成果について考察し、結論を述べる。

第 1 章　音声生成モデルに基く高能率符号化における音源情報抽出とモデル化

1.1　概要

緒論において、高度に発達した現代の情報化社会における音声情報の役割と、音声の高能率符号化に基く情報圧縮が不可欠であることを明らかにした。

本章においてはまず第一に、音声生成モデルに基く音声情報圧縮の妥当性について明らかにし、それに基く代表的な音声符号化方式として LPC 方式について概観する。

LPC 方式においては、残差信号から音源情報をいかに正確に抽出するかが合成音声の音質を大きく左右する。そこで第二に、音源情報の抽出とそのモデル化の必要性について述べ、それに関する従来の諸研究を概観し、その問題点を明かにして本研究への導入を行う。

1.2　音声生成モデルと情報圧縮

音声情報は、緒論において述べたように言語情報と自然性に関する情報、及びそれらの相互作用から成る多様性を有している。そして、これらの情報は音声生成機構なくしては決して発生することができない。従って、音声の符号化における音声信号処理の目的が、音声情報を最大限保存するように表現・変換することであることを考えると、音声生成機構を理解することなしに高能率符号化を達成することは困難である。

一方、人間は発声時、すなわち音声情報の encoding の過程においては、調音運動の統合として音声を発声しており、また、聴取時、すなわち音声情報の decoding の過程においては、音声の複雑な生成機構を熟知した上で複数の音声情報を統合して解読している。そして、上記 encoding 及び decoding の過程においては多くの物理的な冗長が存在する。従って、このような冗長性を効率良く除き高能率な符号化を行うためにも、音声生成機構の理解は必要不可欠である。

以下、音声生成のプロセスについて簡単に述べる [3]。まず音声生成は大きく分けて、音源の発生・調音・及び放射の３つの過程から成る。また、音声の発生器官としては、肺 (lung)、気管 (trachea)、喉頭 (larynx)、咽頭 (pharynx)、鼻腔 (nasal cavity)、口腔 (oral cavity) などが連なっており、一般に喉頭より上部は声道 (vocal tract) と呼ばれている。

音声の発生においては、まず音源が生成される。音源は音声のエネルギー源であり、有声音源 (voiced source) と無声音源 (unvoiced source) とがある。有声音源は、腹筋が横隔膜を押すことにより肺から押し出された空気が、気管から喉頭の声門 (glottis) を通過する際に、声門部の声帯 (vocal cords) が筋肉の弾性による復元力と、空気流によるベルヌイの力との相互作用によって振動することにより発生される準周期的な空気流のパルスであり、声帯音源 (glottal source) とも呼ばれる。この時のパルスの周期は、基本周波数 (fundamental frequency) 又はピッチ周波数 (pitch frequency) と呼ばれ、声帯の緊張度と声門下圧 (subglottal airpressure) に関係する物理量である。ピッチ周波数は、音声の高さ、アクセント、イントネーションなどを決定づける重要な音声情報である。次に無声音源は、舌や口唇などによって形成される狭い空間を空気が通過する際に生ずる乱流によって発生される。このような音源生成過程によって発生された空気流は、顎、舌、口

唇などを動かしてその形を変化させた声道によって共鳴されて音韻が形成され、唇から放射される。有声音源によって発生される音声が有声音 (voiced sound)、無声音源によって発生される音声が無声音 (unvoiced sound) である。

以上述べた音声発生の過程は大まかなものであり、実際には鼻などの影響を受けて複雑に変化するが、この過程は音声発生過程の大部分の特徴を含んでおり、これに基いて音声生成モデルを構築するのは非常に妥当なアプローチであると考えられる。音声発生の過程をモデル化する目的は、簡単かつ有効なモデル化を行うことによって経済的かつ高品質な合成音声を得ることにある。この目的に鑑み、かつ今まで述べてきた音声発生過程を基にすれば、図 1.1 に示すように音源の周波数特性 $G(\omega)$ と調音の周波数特性 $H(\omega)$ を完全に分離し、各々の電気的等価回路が相互作用を伴わずに接続された形で音声波形の周波数特性 $S(\omega)$ が生成されるものとしてモデル化する線形分離等化回路モデル [4] が、最も妥当なモデルとして受け入れられている。

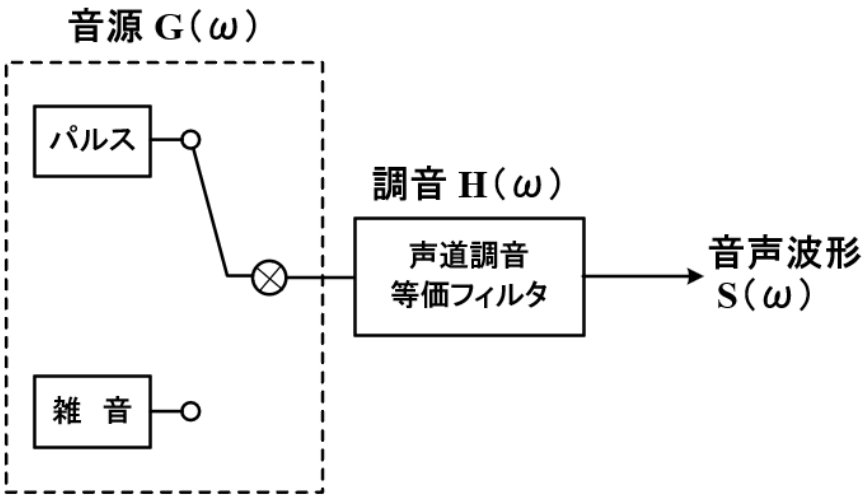


図 1.1 線形分離等価回路モデル

Fig. 1.1 The linear separated equivalent circuit model.

すなわち、音声の周波数特性 $S(\omega)$ は、次の 1.1 式として表現される。

$$S(\omega) = G(\omega)H(\omega) \tag{1.1}$$

この場合、音源は通常、パルス音源と雑音源で近似し、調音は全極型又は極零型のフィルタによって近似する。この時、声帯音源のマクロな周波数特性は、放射特性と共に調音フィルタ特性 $H(\omega)$ に織り込まれる。また、1.1 式における音源特性 $G(\omega)$ は、マクロ的には平坦な周波数特性を有する。

このような生成モデルを導入することにより、音声信号を小数のパラメ－タで正確に記述することが可能となり、そのパラメ－タの時間的変化は音声信号そのものの時間変化に比較して格段に緩やかなものにできる。これにより、音声情報を大幅に圧縮することが可能になる。以上述べた線形分離等価回路モデルに基いて更に具体的なモデルを決定する場合には、以下に示すような条件に留意する必要がある。すなわち、

1. モデルの適合性：過程したモデルが、現実の音声生成過程をどの程度精度良く記述できるか。
2. モデルの効率　：どの程度の数のパラメ－タでそのモデルが既定されるか。

3. モデルの分析性：そのパラメ－タを具体的な音声信号波形のサンプル値から推定できる有効な手段があるか。

などの条件が満たされて初めて最適なモデルを決定することができる。

1.3 線形予測分析 (LPC) 方式

1.2 節で述べた線形分離等価回路モデルに基く音声生成モデルの１つとして極零モデルがある。このモデルは、音声の調音特性を最も正確に表現するモデルであり、最小２乗推定の考え方をもとにした極零モデルの推定法が数々提案されており [5]-[9]、その有効性が確認されている。しかし極零モデルの最小２乗推定においては、方程式の求解が非線形問題となるため反復近似法による最小２乗推定を行う必要がある。また極と零を全く独立に推定することはできず、相互に影響を及ぼし合うため最適な次数の設定も必要となる。そして各処理とも多くの計算量を必要とするため、極零モデルによる音声の分析合成システムの実現を考えた場合、高いコストがかかる。

これに対して、全極モデルに基く線形予測分析がある。線形予測という概念は、Wiener[10] によって初めて導入されたが、音声の分析合成に最初に用いたのは、板倉と斉藤 [11] 及び Atal と Schroeder[12] であり、現在の音声分析合成の基本を成す理論である。

この分析法は、音声波形の標本値間に高い相関があることを利用し、現在の信号 S_n を過去の p 個 (10 個程度) の標本の線形結合である予測値と、その時の誤差信号の和として表そうとするものであり、次の 1.2 式に帰着できる。

$$S_n = \sum_{i=1}^{p} a_i S_{n-1} + \epsilon_n \tag{1.2}$$

ここで a_i は線形予測係数 (LPC 係数)，ϵ_n は予測誤差信号と呼ばれ、この予測誤差信号 ϵ_n の一定区間の平均２乗誤差を最小にするという条件で LPC 係数を求めることを線形予測分析 (Linear Prediction Coding : LPC) という。上記の条件を満たす LPC 係数を求めるためには、原音声信号の一定区間の自己相関関数 r_i を求め、その r_i を用いて、次の 1.3 式で示される正規方程式と呼ばれる p 元連立方程式を解くことによって実現される [19][14]。

$$\sum_{i=1}^{p} a_i r_{k-i} = -r_k \tag{1.3}$$

次に、LPC 分析の周波数域での意味について考えてみる。今、1.2 式の両辺の z 変換をとると、次の 1.4 式となる。

$$S(z) = \{a_i z^{-1} + \cdots + a_p z^{-p}\} S(z) + E(z) \tag{1.4}$$

この式より、次の 1.5 式が得られる。

$$\begin{aligned} S(z) &= E(z)H(z) \\ H(z) &= \frac{1}{A(z)} \\ A(Z) &= 1 - (a_i \mathrm{z}^{-1} + \cdots + a_p \mathrm{z}^{-p}) \end{aligned} \tag{1.5}$$

これより、$E(z)$ を線形システム $H(z)$ に入力した後の出力が $S(z)$ であり、更に $H(z)$ が LPC 係数 a_i で決定される p 次方程式 $A(z) = 0$ の根として与えられる p 個の極による全極モデルとなる

ことを表している。すなわち、音声信号に対して LPC 分析を行うことは、全極モデルに基く音声生成システム仮定したのと等価である。

上記のような全極モデルによる近似を音声に適用した場合、その極は音声スペクトルのホルマントに対応する。そして 1.2 式の予測誤差を最小にするようにして決定された LPC 係数 a_i によって定まる 1.5 式の周波数特性 $H(z)$ は、図 1.1 の線形分離等価回路モデルにおける声道フィルタの周波数スペクトル特性に良く対応している。ここで図 1.1 における音源の周波数スペクトル特性は、LPC 分析における全極モデルでは近似しきれなかった予測誤差信号 ϵ_n(1.2 式) の周波数スペクトル特性 $E(z)$ に対応する。図 1.2 に音声信号の振幅スペクトル特性 $|S(\omega)|$ と、LPC 全極近似フィルタの振幅スペクトル特性 $|H(\omega)|$、及び予測誤差信号 ϵ_n の振幅スペクトル特性 $|E(\omega)|$ を示す。これを見てわかるように、LPC 分析によって求まる $|H(\omega)|$ は音声信号の有声音源のピッチに帰因する調波構造の影響を除去し、声道スペクトルのホルマントによるピークを正確に近似している。そして予測誤差の振幅スペクトル $|E(\omega)|$ は概略平坦な包絡特性を有するが、音源に関する調波構造の成分が明確に残っていることがわかる。

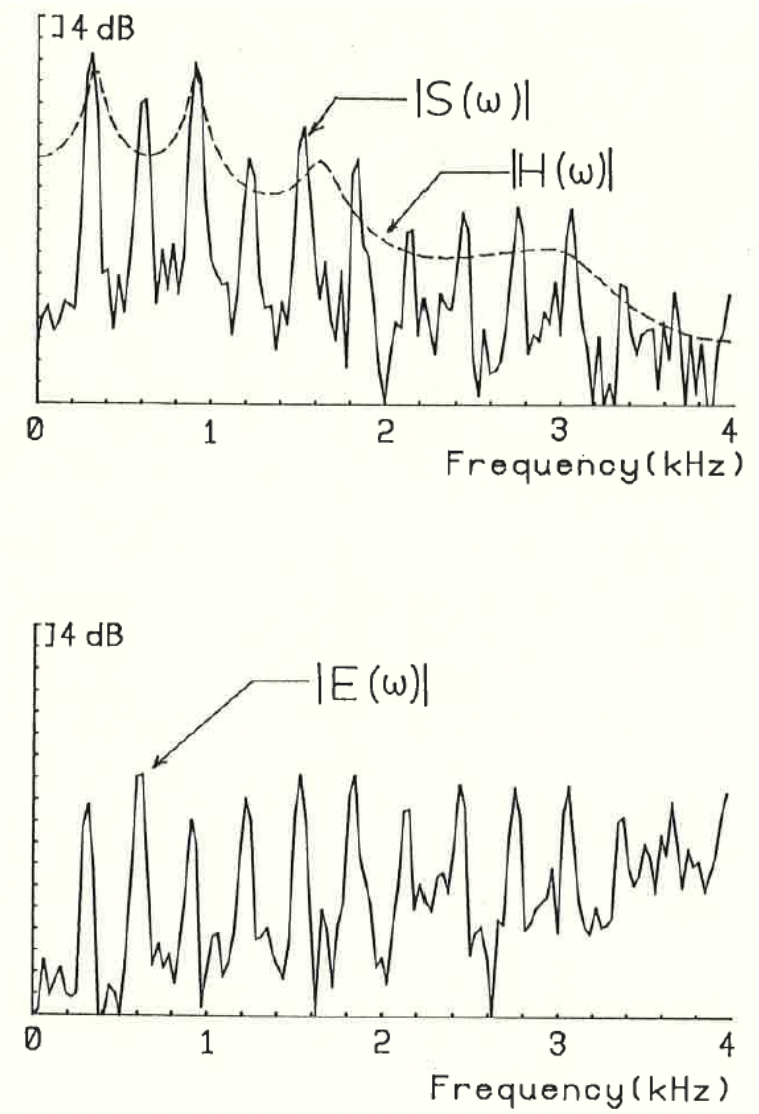


図 1.2 $|S(\omega)|$、$|H(\omega)|$、$|E(\omega)|$ の周波数特性

Fig. 1.2 Frequency characteristics of $|S(\omega)|$,$|H(\omega)|$,$|E(\omega)|$.

1.4 LPC に基く音声合成及び残差からの音源情報抽出とモデル化

以上の理論によって求められる LPC 係数は、音声信号の周波数特性 $S(\omega)$ が定常と考えられる範囲内で、音声信号を全極モデルによる音声生成モデルで近似した時のシステムパラメ－タとなっており、この係数により音声情報としての調音特性を抽出することができ、これらを符号化することにより音声の高能率符号化を実現できる。

ここで 1.3 節の 1.2 式及び 1.5 式は、各々、次の 1.6 式及び 1.7 式と書き直すことができる。

$$\epsilon_n = S_n - \sum_{i=1}^{p} a_i S_{n-i} \tag{1.6}$$

$$E(z) = \frac{S(z)}{H(z)} = S(z)A(z) \tag{1.7}$$

これより予測誤差信号 ϵ_n は図 1.3 に示すように、原音声信号 S_n を LPC 分析によってモデル化された線形システムの逆特性 $A(z)$ を有する逆フィルタに通すことによって得ることができる。

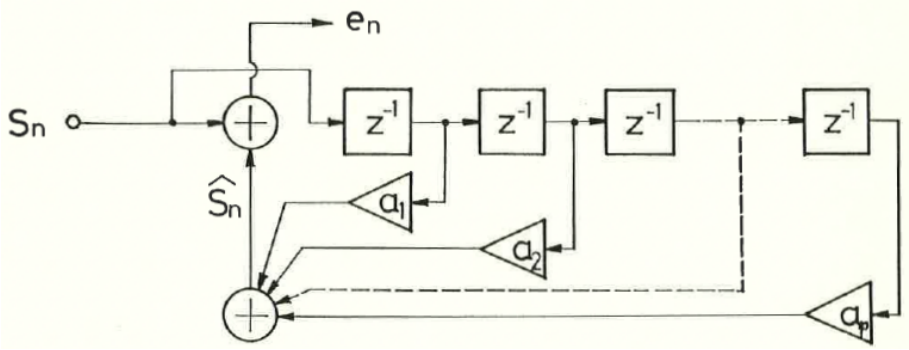


図 1.3 LPC 逆フィルタ

Fig. 1.3 The LPC inverse filter.

上記の信号は残差信号 (residual signal) とも呼ばれ、これは LPC 分析における本来の予測誤差という意味の他に、前記 1.5 式又は図 1.4 のように見れば、出力 S_n を得るための系への入力、即ち音源と考えることができる。

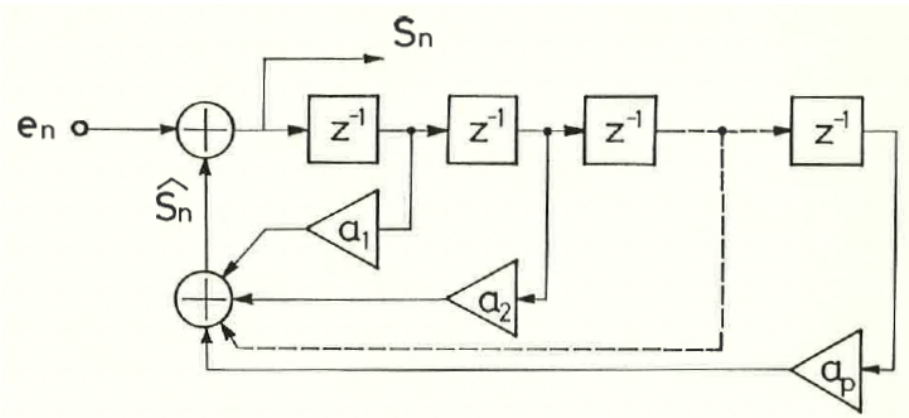


図 1.4 LPC 合成フィルタ

Fig. 1.4 The LPC synthesis filter.

しかし残差信号は複雑な波形信号であり、その情報量は原音声信号とあまり変わらないため、そのまま用いたのでは情報量の圧縮を達成することは難しい。そこで残差信号 ϵ_n を音源の生成過程から考えて妥当と思われる近似の下でモデル化することを考え、まず、その性質について概観してみる。

図 1.5 は、残差信号の時間域の波形の例を元の音声信号と対比させて示したものである。

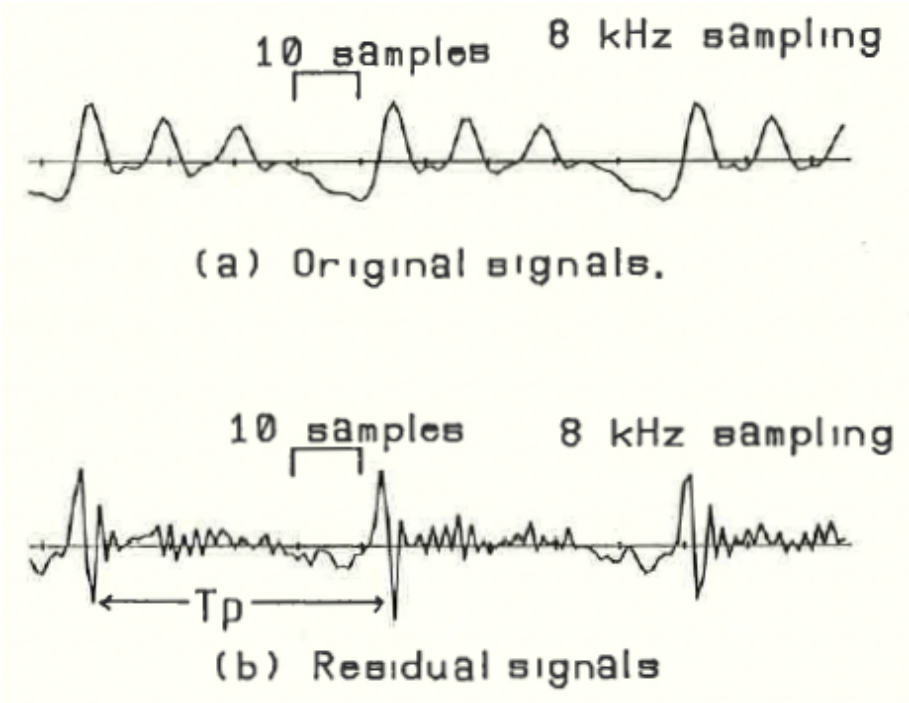


図 1.5 原音声信号と LPC 残差信号

Fig. 1.5 Original signals and LPC residual signals.

LPC 分析において音声信号の予測が完全に行われれば、残差信号は完全なランダム信号となって平坦なパワースペクトルを有する。しかし実際には、特に音声の大部分を占める有声音部分では 1.2 節で述べたように、その音源は周期 T_p の準周期的なパルス列と考えられる。これに対して LPC 分析においては、外から観測できない入力 (音源) を 1 個のインパルス又は白色雑音と仮定してモデル化を行っている。従って、1 個のインパルスによる出力の範囲では、残差信号はランダムな予測誤差と考えられるが、入力として新しいインパルスが印加された時点では、予測は大幅に狂い振幅の大きな残差信号を生ずる。一度入力が印加され、系の出力がそのインパルス入力に対応する応答と考えられる区間に入ると、残差信号は再び本来の予測誤差となり振幅の小さなランダムな信号となる。この変化が、入力として新しい声帯振動によるパルスが加わる毎に繰り返される。その結果、残差信号には図 1.5(b) に示すように、ピッチパルスと同じ周期 T_p でパルス的信号が現れることになる。一方、入力音源が白色雑音で近似できる無性音部分では、残差信号もほぼ白色雑音となる。このように残差信号は、有声音源における周期性及び無声音源における白色雑音性というマクロな音源情報を含んでおり、音源信号の良い近似となっている。このマクロな音源情報に基けば、残差信号の最も簡単かつ有効なモデルとして図 1.6 に示すように、有声音部分をピッチ周期を有するインパルス列で近似し、無声音部分を白色雑音で近似するモデルが考えられる。

このモデルは、声帯振動及び乱流による音源の生成過程を有効に記述しており、1.2 節において述べた良いモデルの条件であるモデルの適合性を十分に満たすものである。また、残差をこのよ

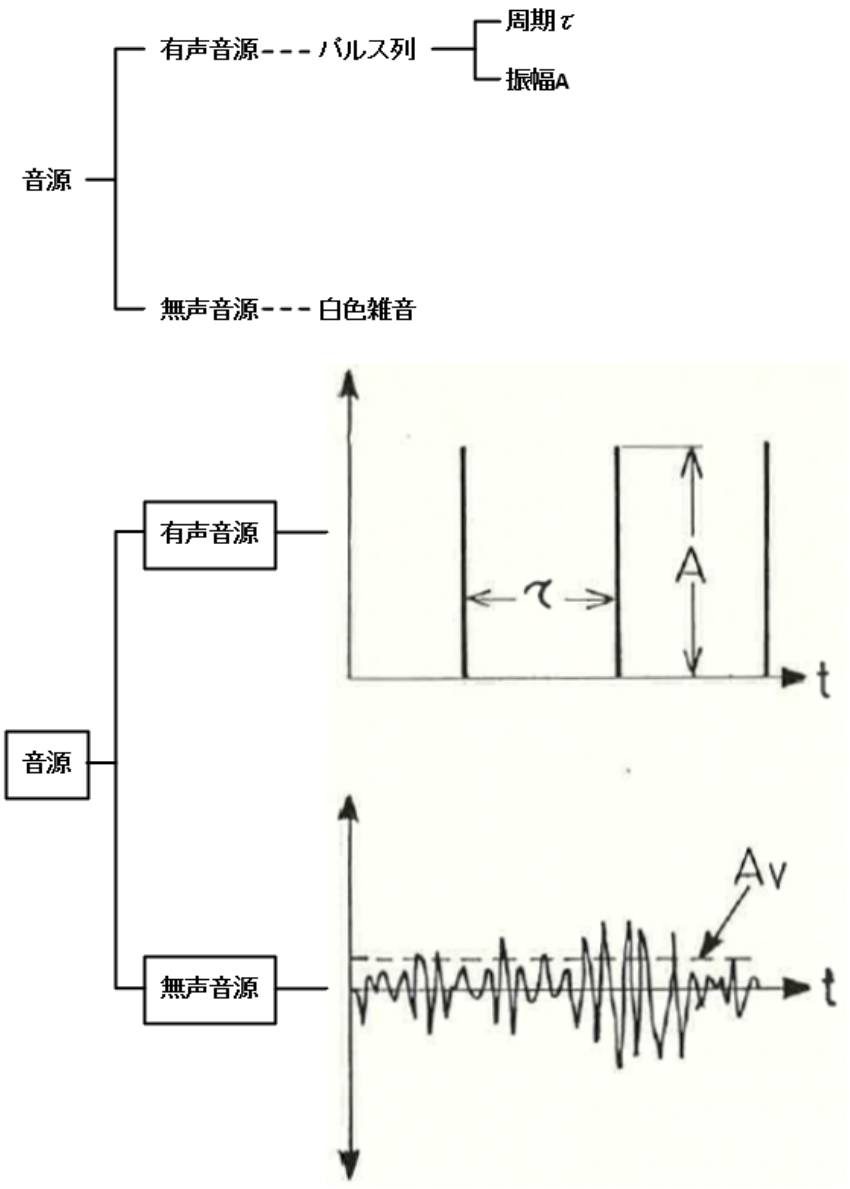


図 1.6 残差信号のマクロなモデル化

Fig. 1.6 Macro-modeling of residual signals.

うにモデル化することにより、一定区間 (数百サンプル) の残差信号をピッチ周期と振幅という 2 つのパラメータに圧縮することが可能であり、良いモデルのもう 1 つの条件であるモデルの効率という面では非常に大きな効果を生む。

次に、残差のミクロな特性について考えてみる。1.3 節の 1.5 式から、音声信号の周波数特性 $S(\mathrm{e}^{j\omega})$ は、次の 1.8 式に示されるように、LPC 全極近似フィルタの周波数特性 $H(\mathrm{e}^{j\omega})$ と予測残差信号の周波数特性 $E(\mathrm{e}^{j\omega})$ との積で表される。

$$S(\mathrm{e}^{j\omega}) = E(\mathrm{e}^{j\omega})H(\mathrm{e}^{j\omega}) \tag{1.8}$$

そして、LPC 分析は短時間の平均残差電力を最小にするような線形予測係数を求める処理であるから、残差電力を E_p は、1.8 式より、次の 1.9 式と表せる (G は利得)。

$$
\begin{aligned}
E_p &= \frac{G^2}{2\pi}\int_{-\pi}^{\pi}|E(\mathrm{e}^{j\omega})|^2\mathrm{d}\omega \\
&= \frac{G^2}{2\pi}\int_{-\pi}^{\pi}\frac{|S(\mathrm{e}^{j\omega})|^2}{|H(\mathrm{e}^{j\omega})|^2}\mathrm{d}\omega
\end{aligned}
\tag{1.9}
$$

この誤差規準より、$|S(\mathrm{e}^{j\omega})|^2 > |H(\mathrm{e}^{j\omega})|^2$ となる周波数領域の方が $|S(\mathrm{e}^{j\omega})|^2 < |H(\mathrm{e}^{j\omega})|^2$ となる領域より大きい重みがかかる [15]。即ち LPC 分析は、音声スペクトルのパワーの大きいホルマント (山) を良く近似する。これは、LPC 分析が全極モデルを仮定しているということを誤差規準の面から表現したもので、図 1.2 からも明かである。しかし、音声スペクトルには声帯の準周期的な開閉動作及び鼻音の調音機構に起因する零点特性も存在し [16][17]、LPC 分析では上記のように、パワーの小さい零点 (谷) はあまり良く近似しない。従って、1.8 式における残差 $\epsilon(n)$ の周波数特性 $E(\mathrm{e}^{j\omega})$ は、その短時間振幅スペクトルにおいて全体的に平坦な周波数特性を有するが、そのほかに特に有声音区間において、前記したように全極モデルとの差としての零点特性を含む。この零点特性は原音声信号の真の音源特性を示すものではないが、残差信号が線形分離等価回路モデルに基く分析・合成システムとしての音源であると考えれば、広い意味でのミクロな音源情報と考えることができ、残差信号、特に有声音残差を全零モデルで近似することにより合成音声の音質を高めることが可能であると考えられる。

本研究は、低・中ビットレートにおける音声生成モデルに基く LPC 分析合成系の合成音声の音質の向上を目指し、LPC 残差信号からのマクロな音源情報抽出法、及び LPC 分析合成系における残差信号の新たなモデル化に基くミクロな音源情報抽出法という、2 つの大きなテーマについて研究を行った結果について述べたものである。

1.5 音源情報抽出とモデル化に関する従来の諸研究と問題点

この節では、1.4 節において述べたマクロな音源情報抽出及びミクロな音源情報抽出に関する従来の諸研究とその問題点について概観する。

有声音源における周期性及び無声音源における白色雑音性という、マクロな音源情報に基く図 1.6 に示した音源モデルは、モデルの適合性及びモデルの効率という 2 つの条件を満たしていることは既に述べた。しかし、良いモデルであるためには 1.2 節において述べたように、モデルの分析性という条件を満たさなければならない。即ちモデル化された音源情報を、実際の残差信号からどのように抽出するかという問題を解決しなければならない。特に、図 1.6 における有声／無声の判別を含めたピッチ周期の抽出は、音声合成技術が発明されて以来の重要な問題であり、従来数々の研究がなされている [18]。

このうち、従来から最も多く検討され有効とされてきた方法として、自己相関関数を用いた方法がある。今、残差信号 ϵ_n について r 次の自己相関関数 R_τ を、Nを分析区間として、次の 1.10 式により計算する。

$$
R_\tau = \sum_{n=0}^{N-1-\tau}\epsilon_n\epsilon_{n+\tau} \tag{1.10}
$$

有声音区間の残差信号は図 1.5(b) に示したようにピッチ周期 T_p に同期して鋭いピークが現れるため、1.10 式によって計算される上記の自己相関関数 R_τ は、図 1.7 に示すように $\tau = T_p$ において最大値を示す。

この原理に基くピッチ抽出の基本手順を以下に示す。

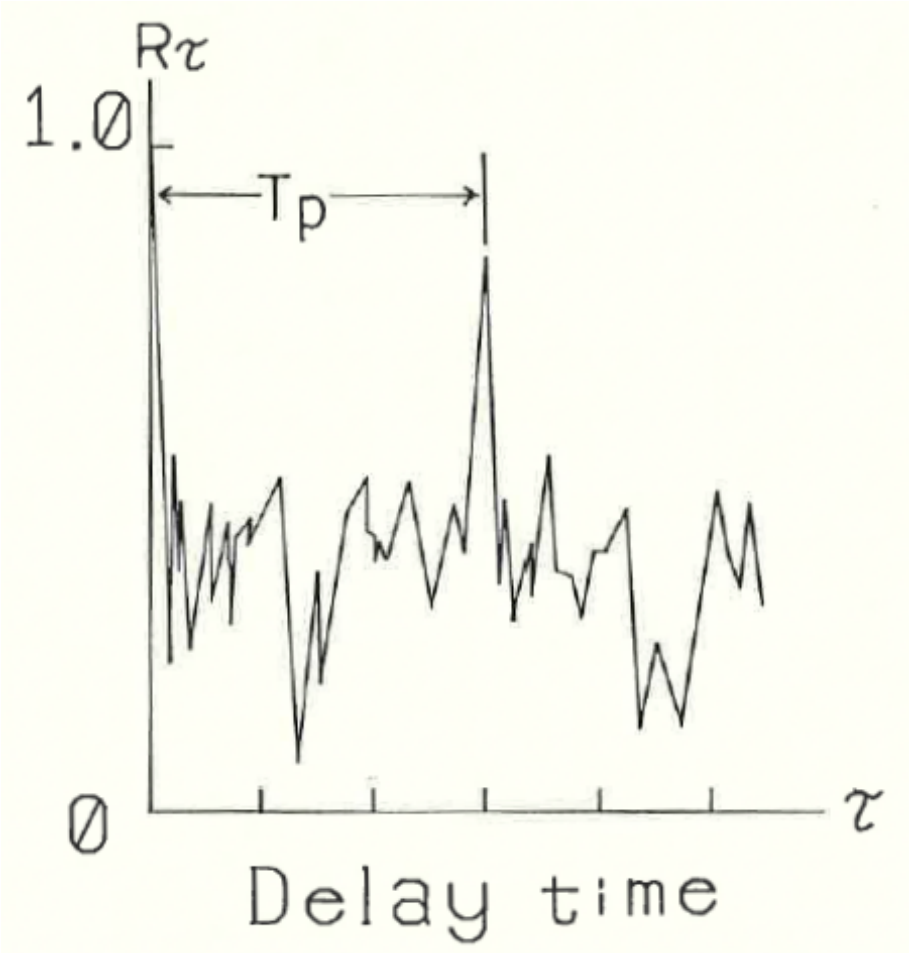


図 1.7 LPC 残差の自己相関関数

Fig. 1.7 The autocorrelation function of LPC residual.

①音源のエネルギーは、残差波形の平均電力 R_0 とする。

②$\tau = 0$ 以外の遅れにおいて最大の相関を与える $\tau = \tau_p$ について R_{τ_p}/R_0 の値 (正規化自己相関) が、0.2〜0.3 以上であればその区間は有声音区間であると判別し、τ_p をピッチ周期として抽出する。それ以下の場合は無声音区間と判別し R_0 を無声音パワーとする。

この理論を応用し実用化された方式として、音声信号を 900Hz 程度の遮断周波数を有するローパスフィルタでろ波した後に LPC 分析を行って得た残差信号に対して、上記のアルゴリズムを適用することにより自己相関関数計算時の時間分解能の不足を補った方式がある [19]。また、1.10 式の積和の代わりに次の 1.11 式に示される、差の絶対値の和で表される平均振幅差関数 (AMDF:Average Magnitude Difference Function) を用いた方式も提案されている [20]。

$$A_\tau = \sum_{n=0}^{N-1-\tau} |\epsilon_n - \epsilon_{n+\tau}| \tag{1.11}$$

この方式では、遅れ τ がピッチ周期に等しい時に平均振幅差関数 A_τ は鋭い谷を示すことによりピッチ周期を抽出でき、乗算がないため計算時間が短くて済むという利点を有する。

この２つのピッチ抽出法は、残差信号に限らず一般の原音声信号にも適用することができるため、原音声信号を３値化 (+1,0,-1) してピッチによる周期性を強調し、かつ３値間の乗算を論理判断のみで行って自己相関をとることにより計算量を減らした方法なども提案されている [21]。

しかし、上記のように残差信号又は原音声信号の相関をとってピッチ周期を抽出する方法は、

低ピッチ話者のときに分析区間の中に入るピッチ周期の数が少なく、十分に大きな値の相関値を得ることができないため抽出誤りを起こしやすい。そしてこの問題に対処するために分析区間を長くとると、ピッチ周波数の変化のために時間分解能が落ち、また高ピッチ話者に対しては分析区間が長すぎてしまうといった相反する問題点を有していた。

相関による方式の他に従来から比較的良く用いられてきた方式として、音声波形からピッチによるピークの候補を複数個選び論理操作によってピッチ周期を決定し、また、このようなピーク検出器を複数個並列に動作させて、多数決によって決定するような方式も古くから提案されている [22]。しかし、このような方式は誤ったピークを抽出することが多いため、結果的に複雑な論理操作が必要になってしまい、抽出率もあまり良くない。

一方、LPC 分析合成系以外の分析合成系を対象とした代表的な方法を参考として挙げておくと、スペクトルを処理することによってピッチ周期を抽出する方法としてケプストラムによるものがある [23]。音声信号をフーリエ変換して得られる周波数特性 $S(\omega)$ は、1.3 節の 1.1 式に示したように調音 $H(\omega)$ と音源 $G(\omega)$ の積で表されるモデルで近似できるため、$S(\omega)$ の対数をとると 1.1 式は次の 1.12 式となって、調音特性と音源特性が分離される。

$$\log S(\omega) = \log H(\omega) + \log G(\omega) \tag{1.12}$$

そしてこの逆フーリエ変換を計算すると、ケプストラム (Cepstrum) と呼ばれる時間域の信号が得られ、有声音区間においては図 1.8 に示すようにピッチ周期に比例した位置に顕著なピークが現れ、これによりピッチ周期を抽出することができる。しかし、この方法においても高ピッチ話者で周波数領域のピッチの倍音数が減るため、ケプストラムにおけるピークが不明確になりピッチ抽出誤差が大きくなるという問題点を有していた。

以上にように従来のピッチ抽出法では、高ピッチ話者と低ピッチ話者の両方でピッチ抽出精度を確保するのが困難であり、また、抽出時の時間分解能は分析する区間長で決ってしまうため、十分な時間分解能が得られないという問題点を有していた。更に、従来の音源情報抽出法においては、音源特性は一定時間内 (20〜30msec) では変化しないという前提に基いてピッチ抽出をおこなっている。しかし、実際にはそのような短時間内においても、無声音区間と有声音区間のわたり部分、及び語尾又は語頭の部分などでは、ピッチ周期及び強度は細かく変動している。従って、このような影響を考慮しない従来のピッチ抽出法では、上記のような細かい変動を平均化してしまうことになり、それにより得られた音源情報に基く合成音声の音質を劣化させてしまうという問題点があった。

残差信号からマクロな音源情報を正確に抽出できれば、図 1.6 に示したマクロな音源モデルに基いて残差信号をモデル化でき、これを音源として LPC 合成を行うことにより実用的な音質を有する合成音声を得ることができる。実際に、マクロな音源モデルを用いた LPC 分析合成方式は PARCOR 方式 [24][25]、LSP 方式 [26] などとして改良され、1.2Kbits/sec〜4.8Kbits/sec の低ビットレートの伝送帯域において十分に高能率な音声符号化を実現してきた。しかし、移動無線等に適用することを目的とした音声の中品質伝送などにおいては、4.8Kbits/sec〜9.6Kbits/sec の低・中ビットレートの音声符号化方式が必要とされ、LPC 方式をその伝送帯域で利用することを考えた場合、4.8Kbits/sec 付近でその音質が飽和してしまい [26]、実用的にはその分野において要求される音質を十分に達成できているとはいえず、その音質の改善が課題となっている。このように音質が飽和してしまう最も大きな要因は、LPC 分析合成方式による音声の調音のモデル化性能が悪いのではなく、残差信号を図 1.6 に示したマクロな音源モデルのみでモデル化したことにある。そこで、残差信号からマクロな音源情報に加えて 1.4 節において述べたミクロな音源情報を抽出

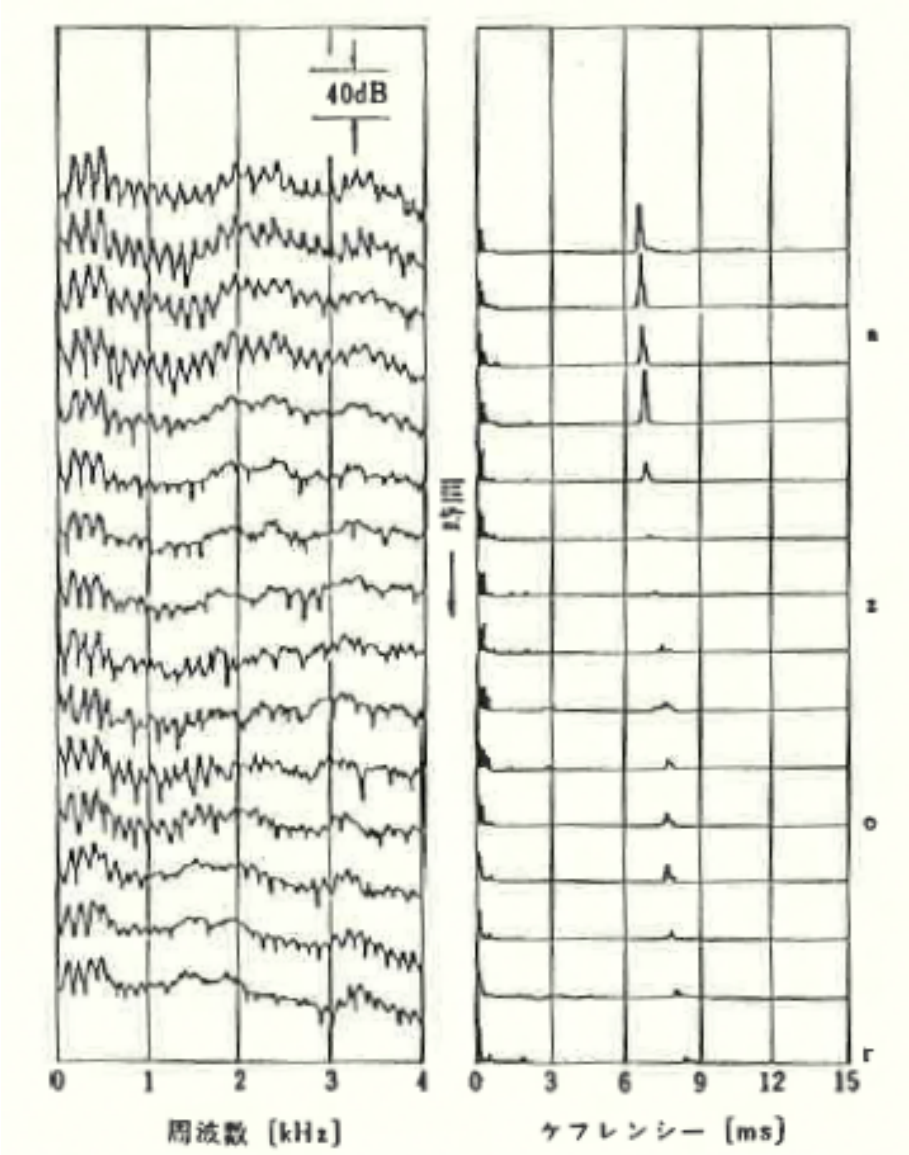


図 1.8 ケプストラム信号

Fig. 1.8 Cepstrum signals.

する意義が生まれてくる。

残差に含まれるミクロな音源情報は、1.4 節において考察したように全極モデルとの差としての零点特性である。この零点特性を残差から抽出するためには、残差の周波数解析を行えばよい。そのための最も一般的な周波数解析法は、短時間離散的フーリエ変換である。これは、残差信号に 20〜30msec の時間窓を乗じたもののフーリエ変換によって定義される。しかしこの解析法は、特に有声音の残差解析を考えたときに 2 つの相反する問題点を有している。第一の問題点として、有声音残差はピッチ周期を有する準周期的信号であるため、時間窓を長くして狭帯域分析を行おうとすると、ピッチ周期で繰り返される時間域のパルス列によって、図 1.2 に示したようにピッチ周期の整数倍の周波数に調波成分が現れてしまい、この調波成分のために零点特性が不明確になってしまう。即ち、ミクロな音源情報がマクロな音源情報にマスクされてしまう。第二の問題点として、上記の調波成分の影響を除くために時間窓を短くして広帯域分析を行おうとすると、窓のフーリエ変換とのたたみこみのために零点特性がぼかされてしまう。このように短時間離散的フーリエ変換は、有声音残差に含まれる調波成分 (マクロな音源情報) 以外の零点特性 (ミクロな音源情報) を精度良く求める分析法としては最適ではない。この問題点の解決法としては、音声

スペクトルの包絡を求める手段として有効なケプストラム分析 [27][28] を用いて、残差の短時間スペクトルの調波成分を平滑化することも考えられるが、元の短時間スペクトルにおいて既に零点特性が調波成分の影響で不明確になってしまっているため、対策としては万全ではない。

以上のように、残差信号のミクロな音源モデルとして零点特性を有する全零モデルというものを予測できながら、その特性を残差信号から取り出す技術がなかった、即ち 1.2 節において述べたモデルの分析性という条件を満たしていなかったため、残差信号にミクロな音源モデルを仮定しそのモデルに基いて残差を符号化した方式は従来なかった。

また、有声音残差の極短時間スペクトルが、ピッチ周期内で高域強調区間と低域強調区間が繰り返して現れるという微細構造を有することを利用し、それに近い図 1.9 に示すようなインパルス列等価音源パルスを有声音源として用いるインパルス列等価音源 [29] を用いた音声合成方式がある。この方式は、残差のミクロな特性に近い音源パルスを用いているという点で有望な方式であるが、残差の分析結果から図 1.9 の形のパルスを最適に決定する手段が示されておらず、モデルの分析性という条件を完全には満たしていなかった。

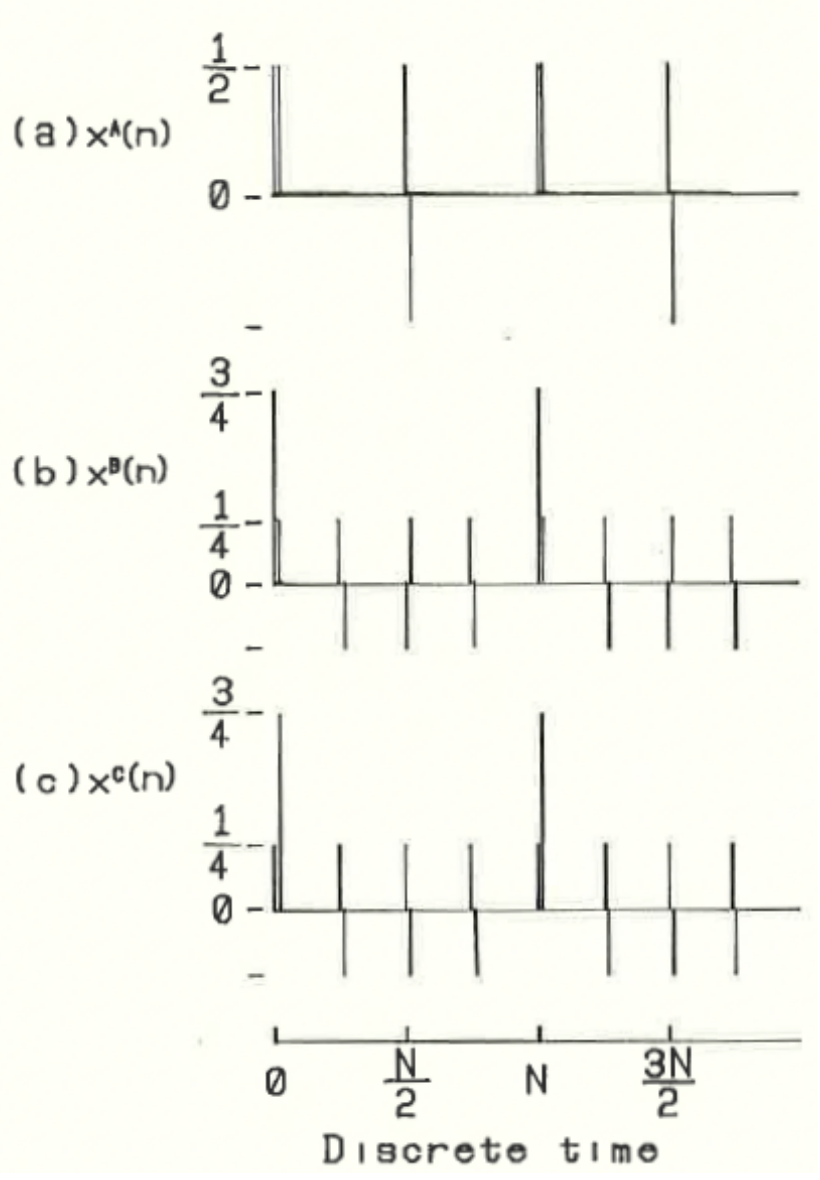


図 1.9 インパルス列等価音源信号

Fig. 1.9 Impulse train equivalent excitation signals.

そこで従来は、残差信号を出来る限り忠実にかつ能率よく符号化するために、残差に生成モデルに基かない適当なモデルを仮定して符号化を行う方式、又は残差だけは波形符号化を行ってその情報を伝送する分析合成符号化と波形符号化のハイブリッド方式などが提案されてきた。これらの従来方式は、残差信号を生成モデルに基いて符号化していないという点で効率的に最適ではなく、また本研究の目指す方向とも異なるが、参考として代表的な幾つかの方式について簡単に挙げておく。

①マルチパルス駆動 LPC 符号化 (MPLPC)[30] 方式

LPC 方式における駆動音源を、1.4 節で述べたような音声生成過程に基くパルスと雑音による音源モデルではなく、複数個のパルスによって近似して LPC 合成フィルタを駆動する方式である。この方式では、基本的には図 1.10 に示すように、Ab-s(Analysis by synthesis) の手法に基く反復演算によって、合成音声と入力音声との誤差を聴覚特性に基く重み付けを行った上で周波数領域で最小とする有限個の駆動パルスの振幅と位置を求め、最適駆動音源パルス情報として伝送し、LPC 合成を行う。

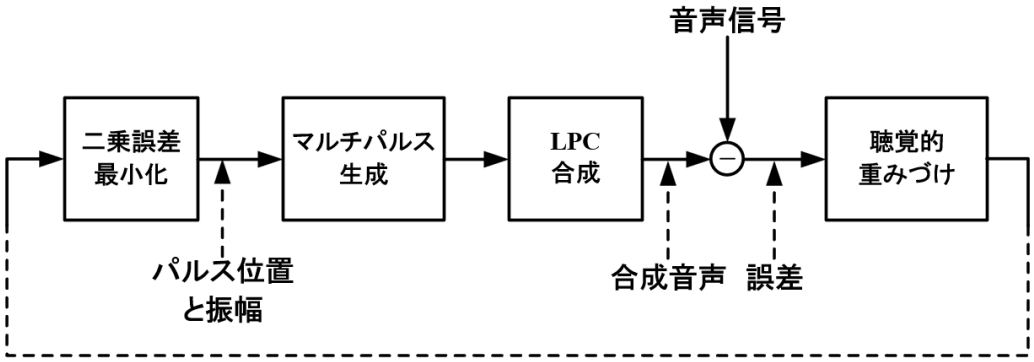


図 1.10 マルチパルス駆動 LPC 符号化

Fig. 1.10 LPC coding by multi-pulse excitation.

②コード駆動 LPC 符号化 (CELP)[31] 方式

LPC 方式においては、予測が完全であれば残差信号は白色雑音となることを利用し、まず、通常の LPC 残差を LPC 逆フィルタによって求めた後、残差に対する長期予測によってパルス列の成分を除いた波形を求め、これを予め用意したコードブック内の所定長の雑音列に当てはめ、この雑音列を音源として LPC 合成フィルタを駆動する方式である。この時の最適な雑音列は、MPLPC 方式の場合と同様に図 1.11 に示すように、雑音列で合成した音声と入力音声との誤差を最小とする雑音列をコードブックから選択してそのコード番号を伝送し、LPC 合成を行う。

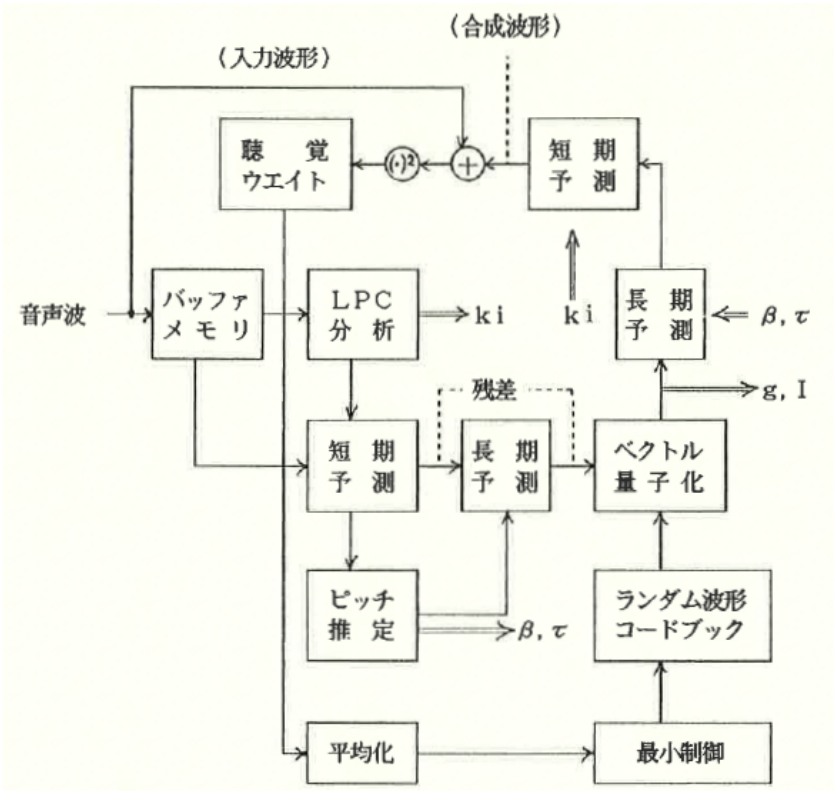


図 1.11 コード駆動 LPC 符号化

Fig. 1.11 The code-excited LPC coding.

③残差駆動 LPC 符号化 (RELP)[32] 方式

図 1.12 に示すように LPC 残差信号の低周波成分、例えば 800 Hz 以下をダウンサンプリングした後に波形符号化して伝送し、受信側では高域周波数成分を整流、クリッピングなどの非線形処理等によって低周波成分から近似的に再生し、LPC 合成を行う方式である。

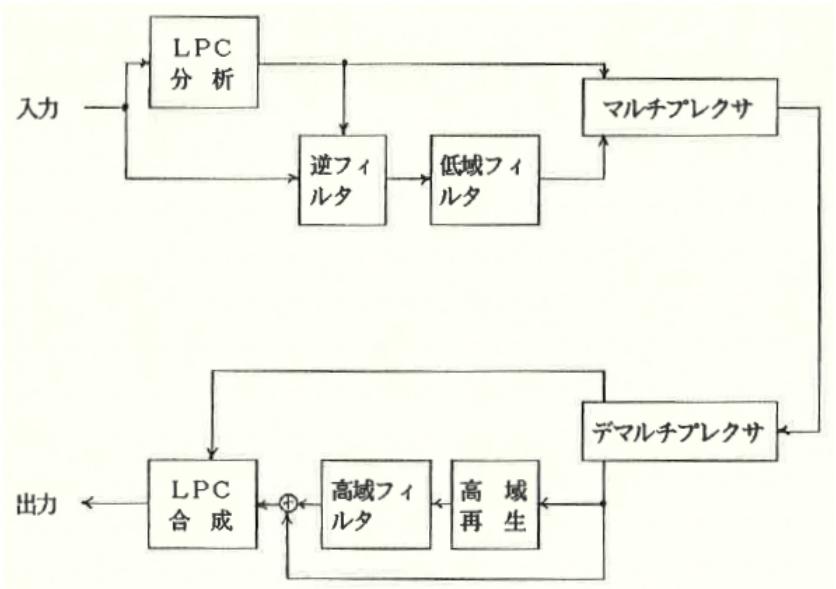


図 1.12 残差駆動 LPC 符号化

Fig. 1.12 The residual-excited LPC coding.

上記の各方式は、低・中ビットレートの音声符号化方式としてかなりの成功をおさめているが、LPC 方式が声道を良くモデル化していることを考えると、その残差、特に有声音区間の残差には、1.4 節において述べたように声帯の準周期的な開閉動作によるミクロな音源特性が含まれると考えるのが自然であり、それに基く音源モデルを仮定できれば更に符号化効率を高めることが可能であると考えられる。

1.6 総括

第 1 章では、まず、音声生成モデルに基く音声情報圧縮の妥当性について明らかにし、線形分離等価回路モデルが音声の高能率な符号化を実現する有効なモデルであることを示した。

次に LPC 分析合成方式について概観し、その性質を明かにした。そして LPC 方式において、LPC 残差信号から音源情報を抽出しモデル化をすることが、高音質な合成音声を得るために必要不可欠であることを示した。特に、マクロな音源情報としてピッチ周期を有するパルス列と白色雑音によるモデル化が有効であり、更にこれに加え、LPC 方式におけるミクロな音源情報として零点特性をモデル化することが必要であることを明らかにした。

そして、マクロ及びミクロな音源情報抽出とモデル化に関する従来の諸研究について概観し、その問題点を明かにして本研究への導入とした。

第 2 章　LPC 分析を用いた残差信号の振幅包絡特性からの音源情報抽出

2.1　概要

第 1 章において、LPC 方式において残差信号から音源情報を抽出することの重要性について明らかにし、残差に現れるピッチピークに基くマクロな音源情報を正確に抽出することが、必要であることを示した。

本章では、LPC 分析が音声信号について、変動の大きな周波数域の波形からホルマントによるピークを適切にモデル化できることに注目し、時間域の残差信号を周波数域のスペクトルとみなし、ピッチピークをホルマントピークに対応させることにより、そのピークを LPC 分析による全極モデルで近似する方法を提案する。

また、鼻音部分においてもピッチの抽出を可能とするために、鼻音化音声を判定しその部分では原音声について上記処理を行う方式を提案する。

上記システムを用いて、ピッチピークの抽出精度及び本手法によって抽出された音源情報を用いて実際の音声を合成し、本手法の有効性を確認する。

2.2　音源情報抽出システム

第 2 章で提案する音源情報抽出システムの構成を図 2.1 に示す [34]。

まず入力した原音声信号 S_1 について、LPC 分析を改良した分析法である PARCOR 分析を行い、それにより求まる PARCOR 係数を用いて逆特性フィルタが構成される [33]。この逆特性フィルタは、入力音声信号の声道スペクトル特性の逆伝達特性を有しており、S_1 をこのフィルタに通すことにより原音声信号に対応する残差信号 S_2 が得られる。次に、このようにして求まった残差信号と原音声信号の平均振幅比、及び残差信号の最大値対平均振幅比が計算され、その結果に基いて原音声信号と残差信号のどちらから音源情報を抽出するかが決定される。このようにして選択された信号に対して 2 乗計算が行われ (S_1 が選択された場合は、S_1 を半波整流してから行う)、得られた信号 S_3 を疑似的な周波数域の振幅スペクトルとみなす。この信号 S_3 はさらに 2 乗され、パワースペクトルとみなされた後偶対称化され、逆 FFT によりそのパワースペクトルに対する自己相関関数 r_i が計算される。続いて自己相関関数 ri を用いて LPC 分析が行われ、LPC 係数 a_i が求まる。この LPC 係数 a_i と先に求まった r_i を用いて FFT による包絡計算を行い、前記 2 乗信号 S_3 に対応する包絡信号 S_4 を得る。更に S_4 に対して平方根が計算され、疑似スペクトルとみなした時間域信号 S_1 または S_2 の包絡信号 S_5 が求まる。この S_5 からピークを抽出し、エラー訂正を行った後、有声音源情報 V としてピッチピークの位置 P_j とその振幅 A_mj を抽出する。ピーク抽出過程でその区間が無音声区間と判断されたら、残差信号 S_2 の平均振幅 A_me を計算し、無声音源情報 UV とする。

以上のシステムが本章で提案する音源情報抽出システムの全体的な構成であり、以下この動作及び理論について詳細に述べる。

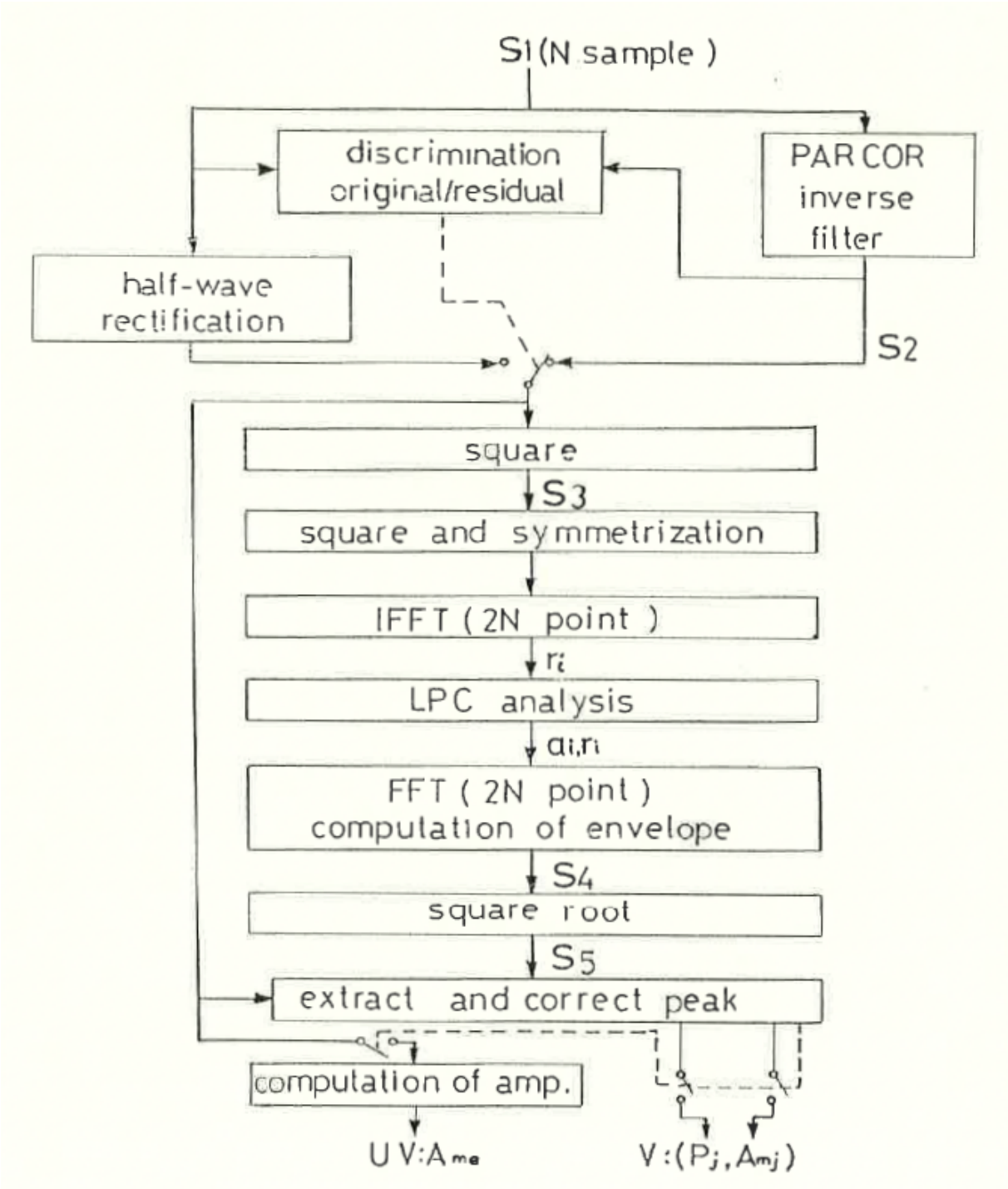


図 2.1 音源情報抽出システム

Fig. 2.1 Block diagram of sound source extraction system.

2.3　本手法における音源情報のモデル化

LPC 分析に基く音声分析合成では、1.4 節で述べたように残差信号をそのまま伝送すれば原音声そのものが再生されるという事実と、同じく有声音における残差信号の予測誤差とピッチパルスとは同期するという事実に基けば、残差信号におけるピークの位置と振幅を図 2 ー 2 に示すように直接モデル化すれば、音源の細かい変動に十分追従できると考えられる。

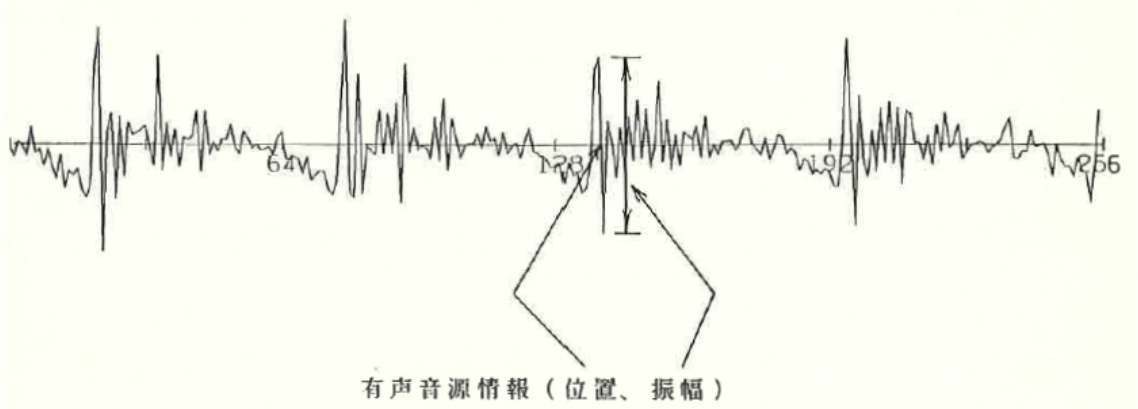


図 2.2 有声音残差の新しいモデル化

Fig. 2.2 New modeling scheme of voiced residual.

しかし、ここで音源のピッチパルスと残差信号の予測誤差の挙動とが、本当に正しく同期しているかという問題がある。今、図 2.3 に示すような音声信号を考えてみる。同図は、「ma」の音声波形とその残差信号である。これを見ると、「ma」が有声音であるにもかかわらず、「m」の部分には残差信号が殆ど現れていない。これは、「m」に代表される鼻音のような音韻が、高調波をほとんど持たない正弦波に近い信号であり、LPC 分析によってそのスペクトル特性が声道特性としてほぼ完全に予測されてしまうためである。従って、このような音声の残差信号から音源情報を抽出した場合、その部分は周期性がなく無声音と判断され、音声合成は白色雑音によって行われることになる。この場合、LPC 係数に「m」のスペクトル情報が保存されているので同じようなスペクトルを有する信号は合成できるが、LPC 係数には位相情報が含まれないため、ピッチ性を表す位相成分が欠落してしまい合成音声が劣化し、また、後続母音「a」の部分とピッチの位相が合わなくなってしまうという問題を生じる。そこで、鼻音部分 (正弦波状音声) についてもピッチが抽出され、その部分の残差信号と同じ振幅を有するピッチインパルス列によって合成が行われれば、ピッチ性を表す位相も正確に再現されると考えられる。

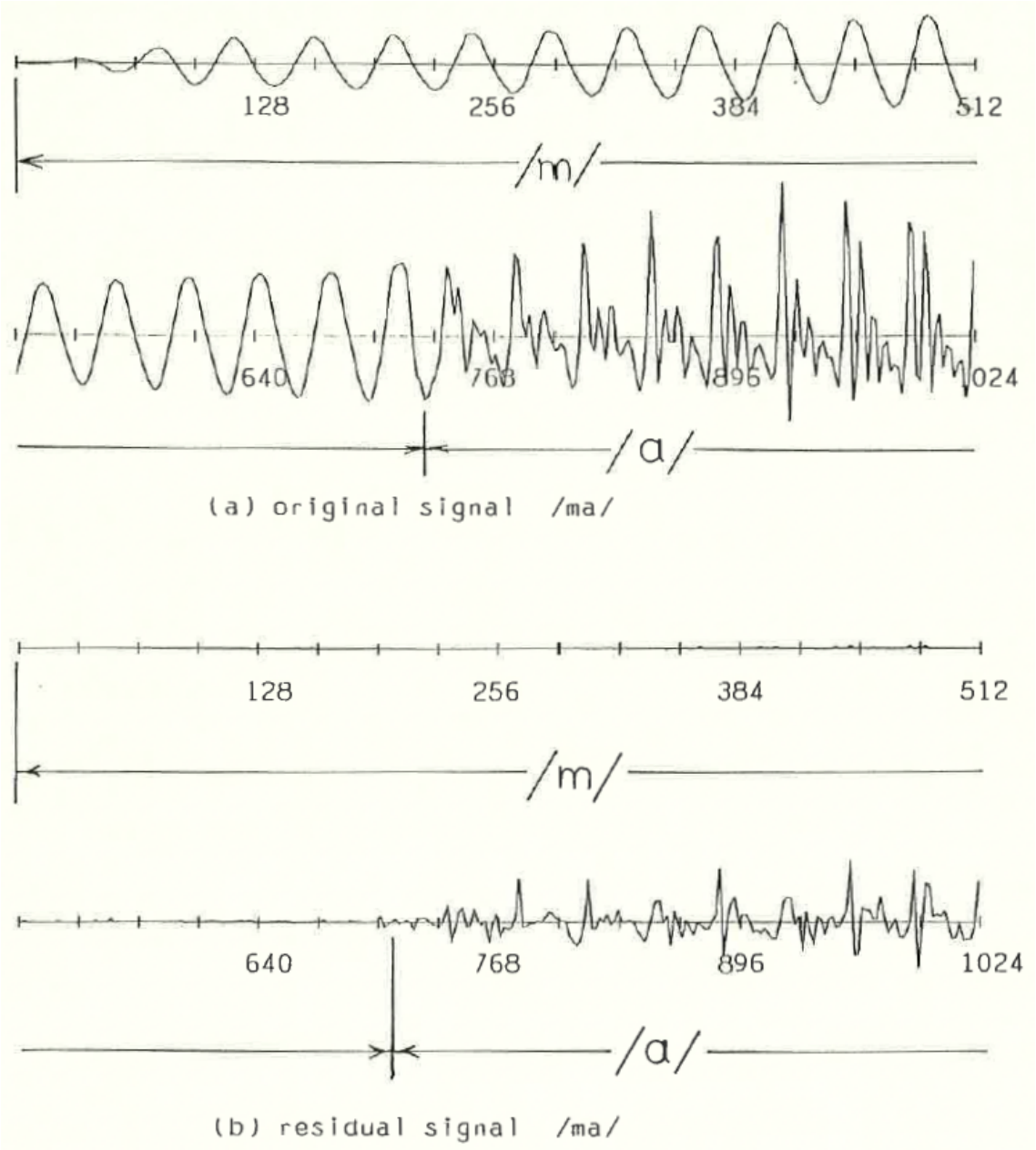


図 2.3 「ma」の音声信号と残差信号

Fig. 2.3 Original and residual signals /ma/.

2.4 残差信号と原音声信号の判定論理

2.3 節で述べたように、原音声信号が鼻音 (「m」など) のように正弦波に近い信号の場合、残差信号からは音源情報を抽出することはできない。従ってそのような音声部分を識別することが必要となる。このような識別を行うための判定論理として 2 種類のパラメータを用いる。まず図 2.4 に、無声音部分と鼻音部分 (「m」) 及び母音部分 (「a」) の音声の原音声信号と残差信号を各々示す。

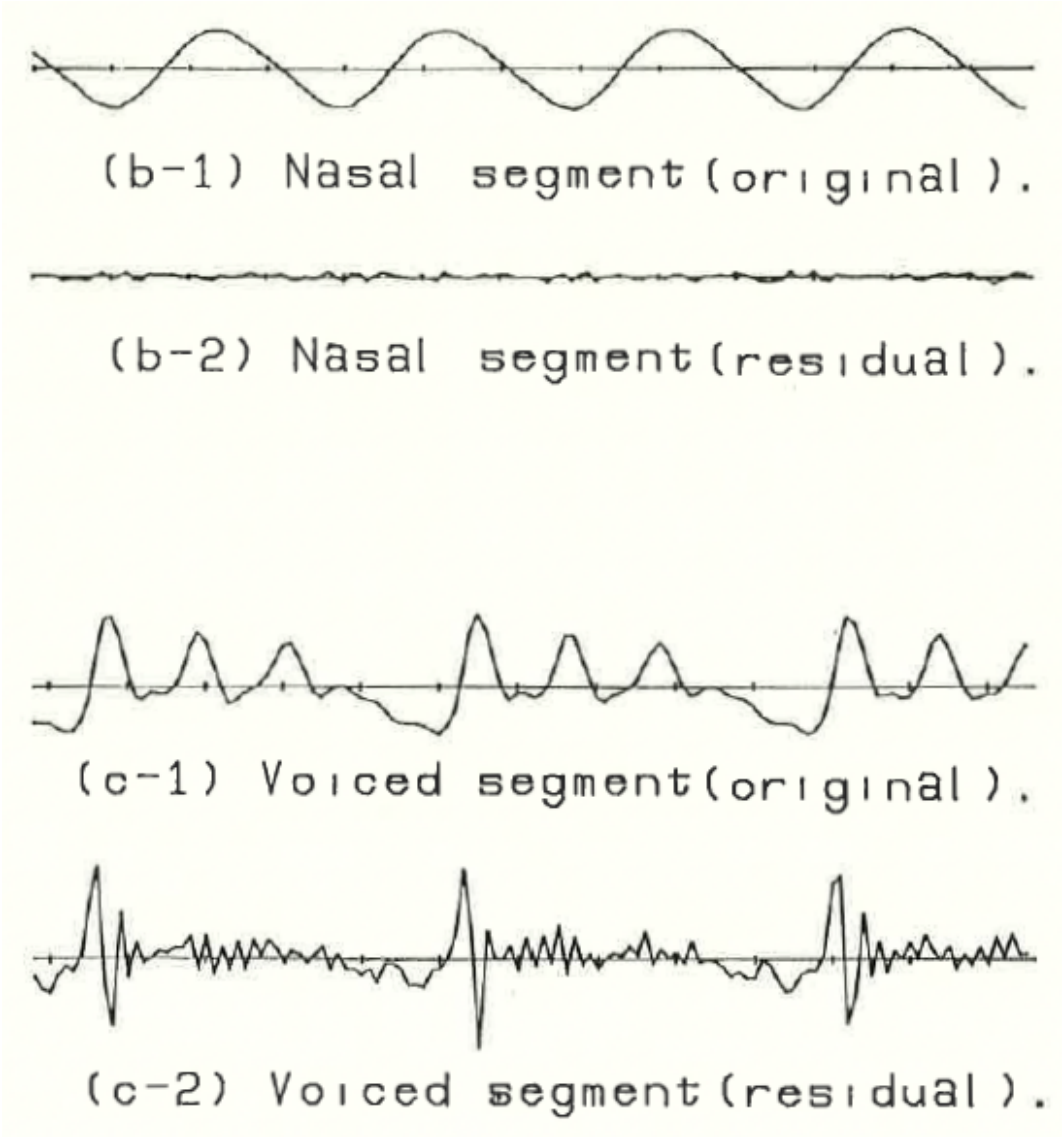


図 2.4 無声音、鼻音、及び母音の音声信号と残差信号

Fig. 2.4 Original signals and residual signals of unvoiced, nasal, and voiced sounds.

これらを見てわかることは、無音声部分と母音部分については、原音声信号と残差信号の振幅の大小関係が互いに一致しているのに対して、鼻音部分については原音声信号のパワーは大きく、残差信号のパワーは小さい。

従って原音声信号のパワーと残差信号のパワーとの比を計算し、そのパワー比が大きければ鼻音部分又は正弦波状音声部分と判定する。ここでは一定区間の平均振幅比 TH_amp を判定論理のための 1 つのパラメータとする。

次に 2 つめのパラメータとして、残差信号における一定区間の最大振幅値とその平均振幅値との比 TH_mav を用いる。この値は同図 2.4 を見れば分かるように、無声音部分及び鼻音部 (正弦波状音声部分) で小さくなる。

以上の 2 つのパラメータ TH_amp 及び TH_mav を用い、一定区間毎の原音声対残差信号の平均振幅比が一定値以上で、かつ残差信号の最大振幅対平均振幅比が一定値以下なら、その区間は正弦波状音声区間と判定し、音源情報抽出を原音声信号から行うこととする。その他の場合は残差信号から音源情報抽出を行うこととする。

2.5 残差信号と原音声信号の振幅包絡特性

2.4 節での判定の結果、分析すべき音声信号が正弦波状音声区間ではないと判定された場合は、残差信号から音源情報を抽出する。

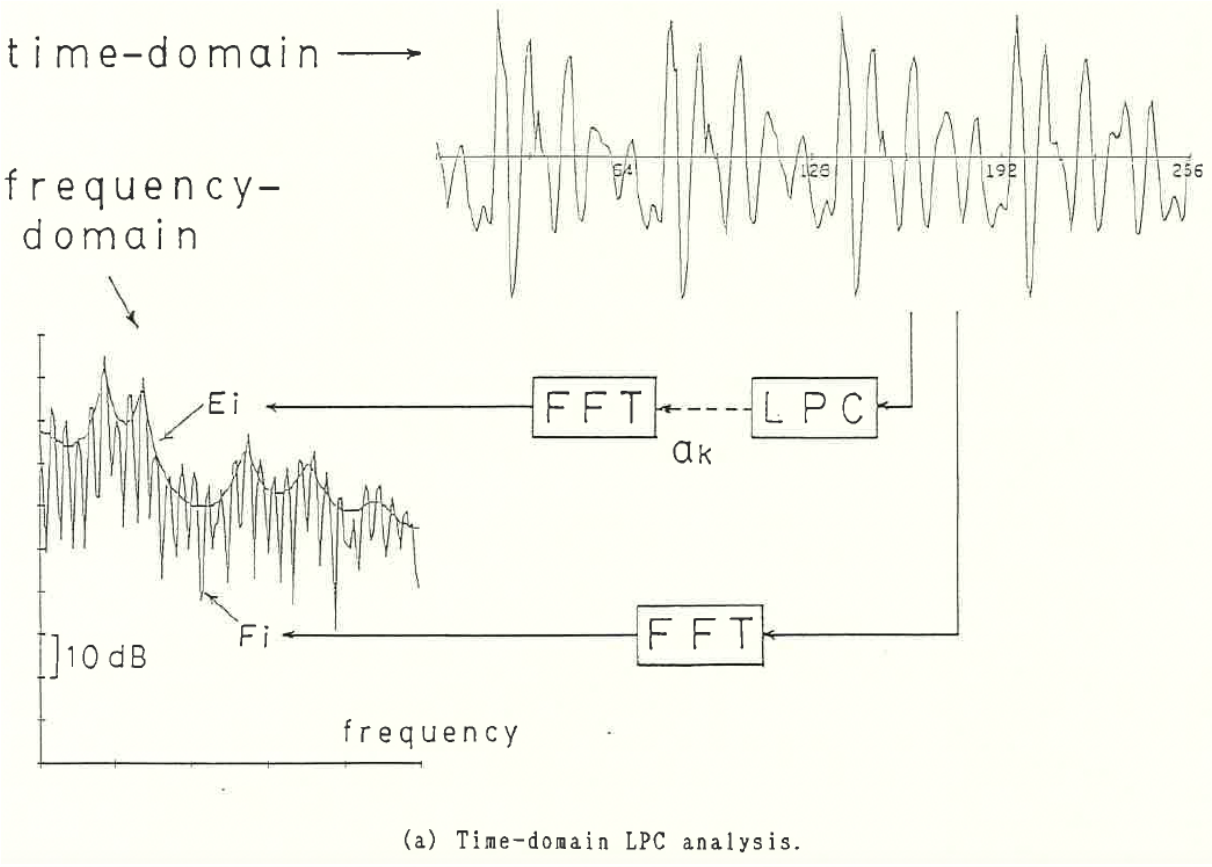
残差信号は図 2.2 に示したように比較的変動の激しい信号である。従ってこの信号からピッチピークの位置と振幅を抽出するために、いかに他の部分の影響を除くかという問題が生じる。本手法ではそのための方法として、LPC 分析を応用したアルゴリズムを提案する [33]。

LPC 分析は 1.3 節で述べたように、時間域音声信号 S_1 の声道スペクトルを全極モデルによって最適に近似する分析法であり、S_1 から求まる自己相関関数 r_i 及び LPC 係数 a_i を用いて、次の 2.1 式を計算することにより、図 2.5(a) のように声道スペクトルの振幅包絡特性 $|H(\mathrm{e}^{j\omega})|$ である E_i を求めることができ、ホルマントによるピークを正確に近似できる [35]。

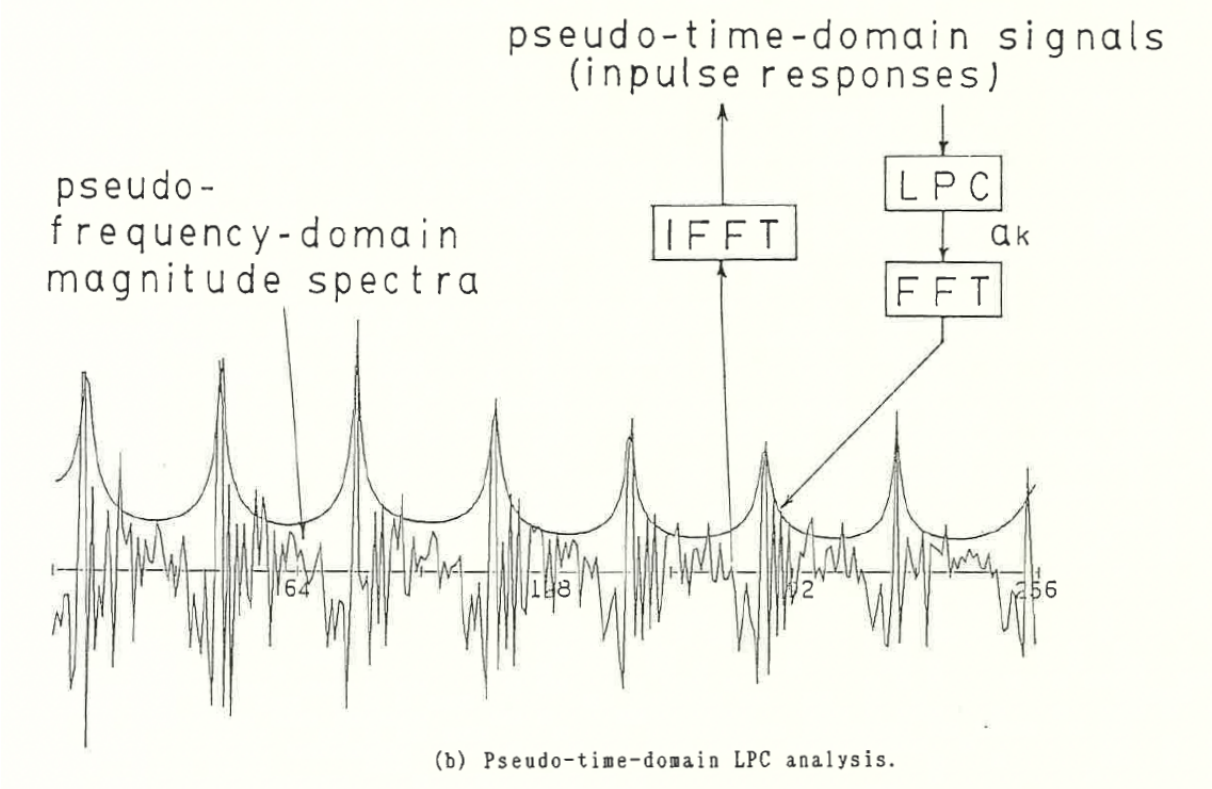
$$
\begin{aligned}
|H(\mathrm{e}^{j\omega})| &= \frac{G}{|1+\sum_{i=1}^{p} a_i \mathrm{e}^{-j\omega_i}|} \\
&= r_0 + \frac{\sum_{i=1}^{p} a_i r_i}{|1+\sum_{i=1}^{p} a_i \mathrm{e}^{-j\omega_i}|}
\end{aligned}
\tag{2.1}
$$

これは言い換えれば、原音声信号 S_1 の周波数振幅スペクトル特性 F_i の変動の激しい成分を除去し、その包絡特性 E_i を求める分析法であるといえる。そこで時間域残差信号 S_2 を、図 2.5(b) に示すように周波数振幅スペクトル特性 Fi に対応させ (これを疑似周波数振幅スペクトルと呼ぶ)、Ei に対応する包絡特性 S_5 を求めることにより、ピッチによるピークをホルマントによるピークと考えて抽出する。以下そのアルゴリズムについて説明する。

上記原理による場合、LPC 分析においては自己相関関数を求めることが必要となるため、上記疑似周波数振幅スペクトルに対応する自己相関関数が必要である。そこで図 2.1 に示したように、N サンプルの入力残差信号 S_2 を 2 乗してピーク部分の変化を強調すると同時に、この信号 S_3 を 1 サンプルめが直流分、N ＋ 1 サンプルめが最高周波数である疑似周波数振幅スペクトルと見なす。この信号 S_3 をさらに 2 乗してパワースペクトルとした後、偶対称化し逆 FFT を行うことによって、上記周波数振幅スペクトル S_3 に対応する自己相関関数 ri を得る [36]。そしてこの ri に対して従来の LPC 分析法を適用して、1．3 節で説明した 1.3 式に従って LPC 係数 ai を求めた後、2.1 式を計算することによって上記疑似周波数スペクトル S_3 の包絡信号 S_4 が求まる。ここで 2.1 式の計算は、FFT を用いて高速に行うことができる [37]。この包絡信号 S_3 の平方根をとれば、時間域残差信号 S_2 の振幅包絡特性 S_5 を求めることができる。



(a) Time-domain LPC analysis.



(b) Pseudo-time-domain LPC analysis.

図 2.5 時間域振幅包絡信号の抽出原理

Fig. 2.5 The principal of extracting time-domain amplitude envelope signals.

以上の処理により、図 2.6 に示すように残差信号 S_2 中の変動の激しい成分を除去し、ピッチピーク部分をよく近似する包絡信号 S_5 を得ることができる。この S_2 と S_5 の関係は前記したように図 2 ー 5 における Fi と Ei の関係によく対応している。

図 2.6 残差信号と振幅包絡信号

Fig. 2.6 Residual signalws and the amplitude envelope signals.

ここで述べた処理は、2.4 節の判定の結果、分析すべき音声信号が正弦波状音声区間であると判定され、原音声信号について音源抽出を行う場合にも全く同様に適用できる。正弦波状音声は図 2.7 の S_1' に示すようにピッチ周期以外の情報をあまり含んでいないため、まずこれを半波整流して隣合うピーク間隔がピッチ間隔になるようにする。この信号に対して上記と全く同じ処理を行うことによって、原音声信号 S_1' の振幅包絡信号を求めることができる。このようにして求まる振幅包絡信号は、前期したように LPC 分析による全極モデルで仮定されているため、図 2.7 の S_5' に示すようにピッチのピーク部分で鋭い極を持つ特性となり、これよりピッチのピークを抽出することができる。

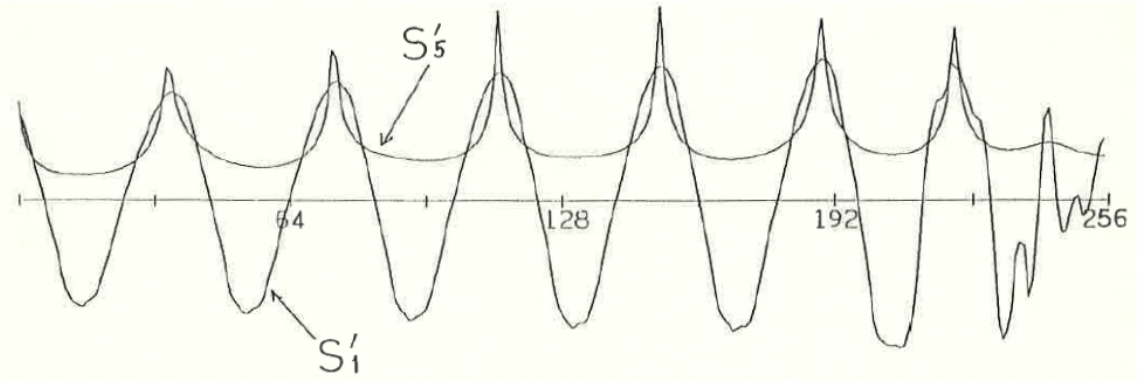


図 2.7 原音声信号と振幅包絡信号

Fig. 2.7 Original signalws and the amplitude envelope signals.

2.6 ピーク抽出・訂正と音源情報抽出

2.5 節で求まった振幅包絡信号 $S_5(S_5')$ は、ピッチピークを極としてモデル化したことにより、それ以外の影響を最小限に抑制した信号である。従って、この $S_5(S_5')$ が極大値をとるサンプル位置と振幅をピッチピークの候補としてのピーク抽出を簡単に行うことができる。

ところで、LPC 分析による全極モデルにおいては、1.3 節の 1.5 式によりモデルを表現する線形システム $H(z)$ が、p 次方程式 $A(z) = 0$ の根として与えられる p 個の極を有することは既に述べた。また、全極モデルによって近似される音声の周波数特性上の 1 つのホルマントは、1 対の複素極によって表現できるため、分析次数はホルマントを表すピーク数の 2 倍程度が最適次数である

[38]。ここで 2.2 節で説明した図 2.1 のシステムにおいて、疑似周波数振幅スペクトルと見なした時間域残差信号 S_2 には、10KHz 標本化音声の場合で分析区間のサンプル数 N が 256 点の場合、ピッチ周波数は最高 500Hz 程度であるため最高 12 個程度のピッチピークが現れる。従って、このシステムにおける LPC の分析次数は 24 次としている。このようにして振幅包絡特性が求まりピーク抽出を行った場合、ピッチ周波数が女性の有声音の場合のように高い場合は、1 分析区間内のピーク数と分析次数とが適切に対応しているため、正しいピークが求まっている。ところがピッチ周波数が男性の有声音の場合のように低い場合は、1 分析区間内のピーク数が少なく分析次数が高すぎるため、真のピークの他に疑似のピークが現れる場合がある。また、無音声と有音声のわたり部分においてもそのようになる傾向がある。

以上のような場合における疑似ピークの除去を行うための訂正アルゴリズムについて以下に例をあげながら説明を行う。訂正アルゴリズムは大きく分けて振幅判別ルーチンと間隔判別ルーチンの 2 種類から構成される。

I. 振幅判別ルーチン

(i)まず振幅が小さく明らかにピッチによるピークでないピークは、値の小さな振幅値 TH_{cut} によって除去される。

(ii)現在判別しようとするピークの振幅 Ap と、その直前の谷の振幅 Av との比を、次の 2.2 式によって計算する。

$$r_p = \frac{A_p}{E_v} \tag{2.2}$$

そして、この値 r_p が一定の値 TH_r 以下なら疑似のピークとして除去する。なおこの判別ルーチンは、1 つ前のピーク位置との間隔が 1 分析区間 (256 サンプル) 以上の場合、又はそれに続く真のピークが数個 (4 個程度) しか求まっていない場合、又は後述する間隔判別ルーチンで判別できなかった場合に働く。

以上の判別ルーチンによって図 2.8(a) のように比較的残差パワーの大きな無声音の残差信号の場合、その包絡特性から波線矢印のようにピークが求まるが、各ピークとも 2.2 式によって計算される r_p の値が閾値以下になるため、全て疑似ピークとして除去される。また同図 (b) のように、有声音の始まり部分のピークを正確に抽出できる。

Ⅱ. 間隔判別ルーチン

前記Ⅰ－ (ii) の振幅判別ルーチンによって、有声音の始まりの数個のピッチピークが抽出された後、現在判別しようとするピークの直前の数ピッチ間隔 (4〜8 間隔) の平均値を計算する。そして現在のピークと 1 つ前のピークとの間隔が平均値と比べて一定誤差以内 (平均値の 10〜20 % 以内) ならば現在のピークは真のピークであるとする。

そうでない場合において、現在の次のピークと 1 つ前の真のピークとの間隔が、平均値と比べて上記一定誤差以内ならば、現在のピークは疑似ピークであるとして除去する。

上記 2 つの場合に当てはまらないときは、判別不能として前記Ⅰ－ (ii) の振幅判別ルーチンによる判別を行う。さらにこのようにして真のピークとして求まったピーク位置との間隔が、前記平均値に比べて 2 倍以上であったら、その部分で音声が途切れたと判断し、その次のピークはⅠ－ (ii) の振幅判別ルーチンのみで抽出を行う。また上記ピーク間隔は次の平均値の計算には含ませないようにする。

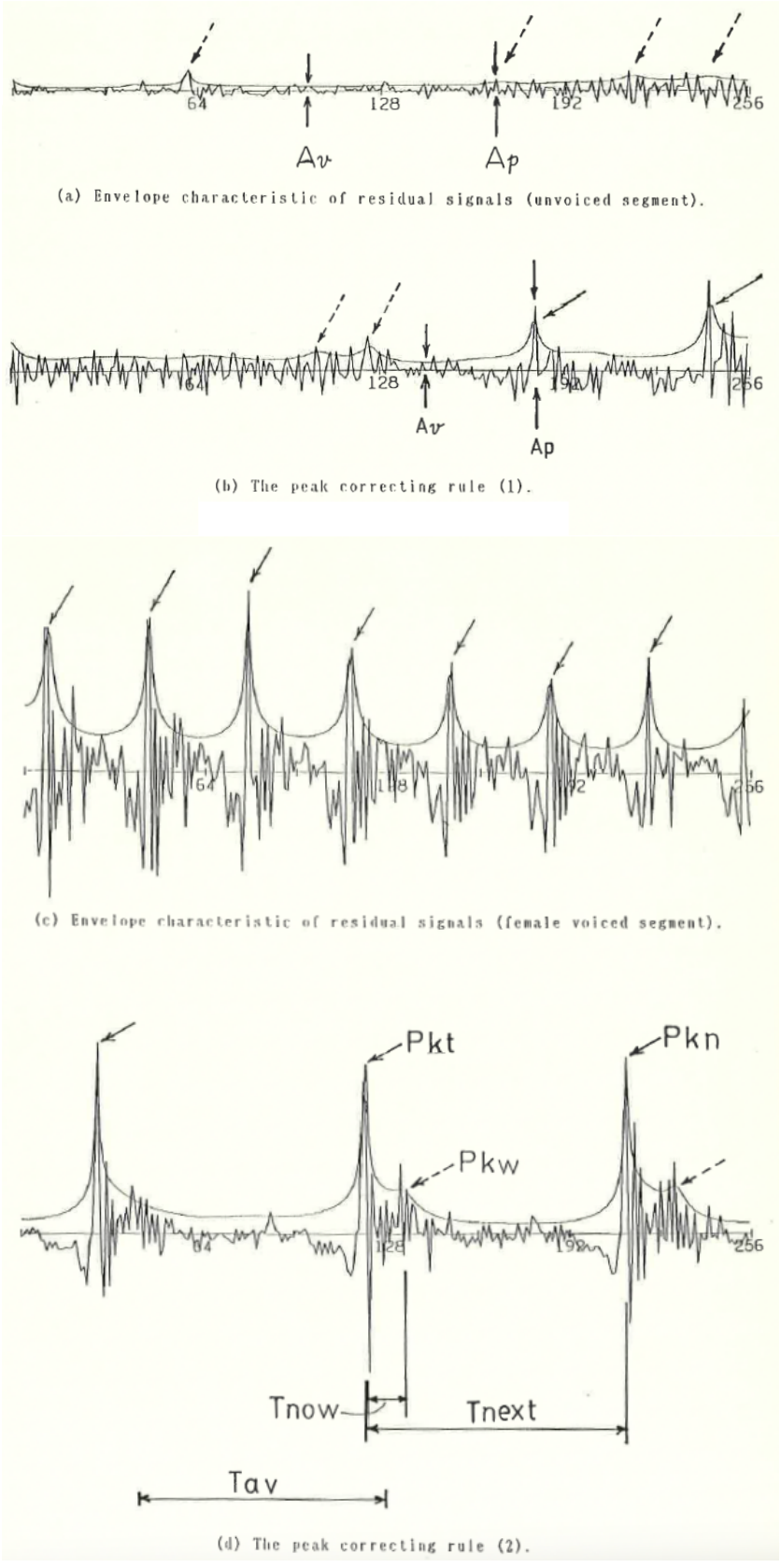


図 2.8 ピーク抽出とピーク訂正

Fig. 2.8 Peak extraction and peak correction.

以上の間隔判別ルーチンは、真のピークの間に疑似ピークが現れるとしても、ピッチピークを全極モデルで近似したときの次数の制約によって、高々 1 つしか現れないと言う事実に基づいている。

この判別ルーチンによって、図 2.8(c) の女性の残差信号のように真のピーク数が多ければ、全て真のピークと判別される。また同図 (d) の男性の残差信号の場合のように真のピーク数 (実線矢印) が少ない場合においては、例えば図中の現在注目しているピーク $P\mathrm{K}_w$ を見た時、$P\mathrm{K}_w$ の位置とその直前の真のピーク $P\mathrm{K}_t$ との間隔 T_{now} と、現在の次のピーク $P\mathrm{K}_n$ と $P\mathrm{K}_t$ との間隔 T_{next} とを比べた場合、T_{next} の方がその直前数ピッチの平均間隔 T_{av} により近い。よって現在のピーク $P\mathrm{K}_w$ は疑似ピークであるとして除去される。またもし間隔判別ルーチンによって判別できなかった場合でも、I − (ii) の振幅判別ルーチンによって計算される r_p の値が値より小さいため除去できる。またピーク $P\mathrm{K}_{next}$ を判別するときに、間隔判別ルーチンによって判別できなかった場合は、同図のようにそれに続く I − (ii) の振幅判別ルーチンにおいて計算される r_p の値が、閾値よりも大きくなるため真のピークと判別される。

上記判別ルーチンによる判別結果に基づいて、次のように音源情報を抽出する。

(Ⅰ) 上記判別ルーチンによって、一分析区間 (256 サンプル) 内で 1 つもピークが抽出されなかった場合は、その区間は無音声区間であるとしてその区間の残差信号の平均振幅 A_{me} を計算し、それを無声音源情報 UV として出力する (図 2.1)。

(Ⅱ) 2.4 節の判別の結果、残差信号について音源情報抽出が行われた場合において、上記の判別ルーチンによって真のピークが抽出された場合、抽出されたピークの位置 $P_j(1 \leqq P_j \leqq 256)$ と、振幅 A_{m_j} を有声音源情報 V として出力する (図 2.1)。)。

(Ⅲ) 2.4 節の判別の結果、原音声信号について音源情報抽出が行われた場合において、上記判別ルーチンによって真のピークが抽出された場合、抽出されたピークの位置 P_j と、その原音声信号に対応する残差信号における位置 P_j 付近の平均振幅 (10〜20 サンプル程度) を、抽出ピークの振幅 A_{m_j} として有声音源情報 V を出力する (図 2.1)。)。

以上のようにして 2.3 節で述べた音源モデルに基づく音源情報が抽出される。以下、図 2.1 に示した本手法による音源情報抽出システムについて実験を行い、その結果について検討を行う。

2.7 実験結果

2.7.1 音源情報抽出の実験条件

実験にはメインシステムとして PDP − 11 を用い、それに 12bit 精度の A/D、D/A が付く。音声データはオープンリールデッキから 5KHz のローパスフィルタを介して、10KHz サンプリングで A/D を行う。またマイクロコンピュータ IF800 を PDP − 11 のグラフィックディスプレイとして用い、同時に XY プロッタが付く。また D/A の出力には 5KHz のローパスフィルタ、アンプ、スピーカが接続される。

実験に用いた音声は 10KHz でサンプリングされ、12bit 精度で量子化されている。そして 20Kword(20480 サンプル)(約 2 秒分) を 1 つの音声データファイルとし、全ての処理を行う。

まず本手法による音源情報抽出システム (図 2.1) を用いて、実際の音声信号から音源情報を抽出した。図 2.1 のシステムにおいて、まず PARCOR 分析による逆特性フィルタは、PARCOR 格子型フィルタを用いて残差化を行う。このとき、PARCOR 分析は 256 サンプルを 1 分析フレーム

とし、分析次数は 10 次とする。また PARCOR 係数は Durbin-Levinson-Itakura 法 [14] によって求める。

次に 2.4 節で説明した残差信号と原音声信号の判定論理、及び 2.5 節の振幅包絡特性を求めるアルゴリズムは、128 サンプルを 1 フレームとし、フレームの両端に 6 サンプルづつの重なりを持たせ、116 サンプルづつデータを更新することによって端の影響を除いている。

次に図 2.9 は図 2.1 における 2 乗 (パワースペクトル化)、偶対称化及び逆 FFT の部分の具体的な処理アルゴリズムである。d_1 がパワースペクトルとした後の 2 乗残差信号、d_2 は逆 FFT への実数部入力信号であり、虚数部入力は全て 0 とする。逆 FFT の結果、実数部出力として自己相関関数 r_i が求まる。なお、図 2.1 の LPC 分析にも、Durbin-Levinson-Itakura 法を用いている。次数は 2.6 節で述べたようにピーク数の 2 倍程度が最適次数であり、1 分析フレーム (128 サンプル) 内に最高 6 個程度 (500Hz のピッチに相当) 現れることを仮定し、12 次とする。従って図 2.9 における自己相関関数は $r_0 - r_{12}$ を用いる。

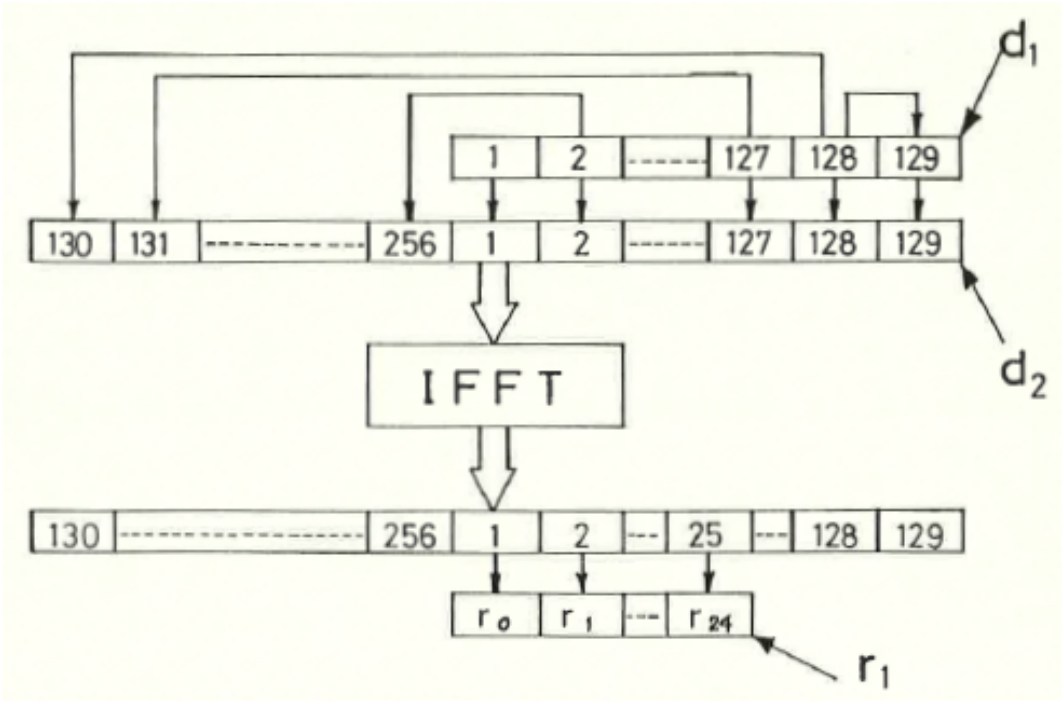


図 2.9 データの偶対称化

Fig. 2.9 Data symmetrization.

図 2.10 は、図 2.1 における FFT 及び包絡計算の部分の具体的な処理アルゴリズムである。d_3 は LPC 分析によって求まる LPC 係数 $a_0 - a_{12}$、d_4 は FFT への実数部入力信号であり、虚数部入力は全て 0 とする。FFT の結果得られる各々 256 個の実数部出力 X_i 及び虚数部出力 Y_i のうち、前半部分 128 個について処理 (a) において 2 乗和の計算が行われ、求まった信号 $|A_i|^2$ について処理 (b) を計算することによって、前記 2.1 式の計算が行われ、図 2.1 における 128 サンプルの 2 乗残差信号 S_3 の包絡信号 S_4 が求まる。この包絡信号 S_4 についてさらに図 2.1 の平方根を取れば、原残差信号の包絡信号 S_5 を求めることができる。

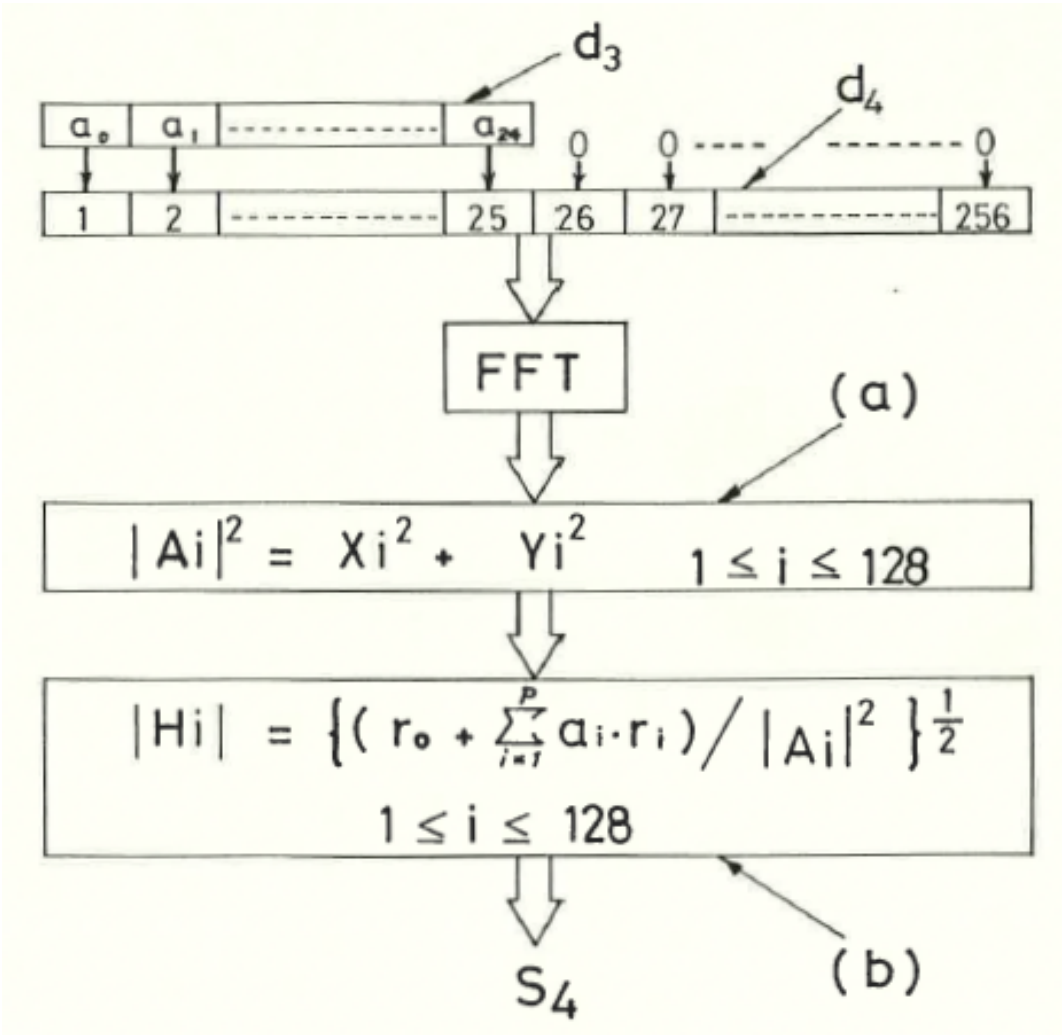


図 2.10 包絡計算

Fig. 2.10 Envelope computation.

以上から求めた S_5 の中央部 116 サンプルについて、極大値を取るサンプルの位置と振幅値からピークを抽出することが出来るが、図 2.10 の処理 (b) における計算式の分子は 1 分析区間において定数となるので、実際のピーク抽出においては、処理 (a) の出力信号 $|A_i|^2$ における極小値をとるサンプルの位置からまずピークの位置が抽出される。そしてピーク位置のサンプルについてのみ、処理 (b) 及び平方根 (図 2.1) の計算が行われ、そのピーク位置における振幅が計算される。このようにすることによって、包絡信号の全てについて処理 (b) 及び平方根の計算をする必要はなくなり、計算量を減らすことが出来る。

以上の処理に続くピーク訂正処理 (2.6 節) は上記振幅包絡計算における分析フレーム長とは異なり、256 サンプルを 1 分析 (訂正) フレームとする。そして有声音の場合、そのピーク位置が 256 サンプル中の位置として抽出され、無声音の場合は 256 サンプルの平均振幅が抽出される。上記振幅包絡計算の分析フレーム長を 256 サンプルにしなかった理由は、包絡計算における FFT、逆 FFT、LPC 分析のための計算回数が、後述するように 256 サンプルについて 1 回の処理よりも、128 サンプルについて 2 回の処理の方が少ないためである。ただし、あまり分析フレーム長を短くすると、分析精度 (ピーク抽出精度) が悪くなるため、予備実験の結果、分析精度が悪くならないフレーム長で 128 サンプルを選択した。

以下上記条件における各実験結果について述べていく。なお実験に用いた音声は以下の男女 5 ファイルずつ、合計 10 ファイルである。

No.1(DATA07.DAT)	「6 月は秋ですか」	(男性)
No.2(DATA07.DAT)	「 〃 」	(女性)
No.3(DATA18.DAT)	「クリスマスは何月にありますか」	(男性)
No.4(DATA18.DAT)	「 〃 」	(女性)
No.5(DATA28.DAT)	「横浜は何県にありますか」	(男性)
No.6(DATA28.DAT)	「 〃 」	(女性)
No.7(XVOI01.DAT)	「明日は雨でしょう」	(男性)
No.8(XVOI03.DAT)	「 〃 」	(女性)
No.9(XVOI02.DAT)	「スキーは南国でできますか」	(男性)
No.10(XVOI04.DAT)	「 〃 」	(女性)

なお各ファイルとも 256 サンプルを 1 ブロック (bl k) として 0 ー 79 ブロックある。

2.7.2 残差信号と原音声信号の判定

まず、2.4 節の残差信号と原音声信号の判定論理のための値を求める予備実験を行った。その結果、原音声対残差信号の平均振幅比の値 TH_{amp}、及び残差信号の最大振幅対平均振幅比の値 TH_{mav} の最適値が、

$TH_{amp} = 7.0$

$TH_{mav} = 6.0$

と求まった。次に、この値を用いて 10 個の音声ファイルにつき、分析フレーム 116 サンプルとして判定を行った。その判定結果の一例を図 2.11(a)、(b)、(c) に示す。図中で縦線で区切られた部分が 1 分析フレーム (116 サンプル) で、× 印のついているフレームが正弦波状音声区間と判断されたフレームである。同図 (a) 及び (b) の × 印となっているフレームは、原音声信号が正弦波状に近いためその残差信号の周期性が不明確になっておりまたパワーも小さいため、原音声信号への切り換えの判定が非常にうまく行われている。また、同図 (c) の例では、× 印となっているフレームにおいても、その残差信号に多少周期性がみられる。しかしこの程度のものなら、原音声信号について音源抽出を行っても正確に求めることができ、後述の抽出精度には殆ど影響を及ぼしていない。以上のアルゴリズムによって、残差信号に周期性が殆ど出ない正弦波状音声区間の判定が、正確に行われることがわかった。

閾値 TH_{amp}、TH_{mav} については 7．0 及び 6．0 が最適値として求まったが、これは通常の有声母音の原音声信号と残差信号の平均振幅比は 7 倍以下であり、また残差信号のピッチによるピークと平均振幅との比が 6 倍以上あれば、周期性があると判断していることになる。そしてその仮定は図 2.11 の結果を見ると、確かに正しいということがわかる。ただしこの方法は非常に簡単な方法であるがゆえに、同図 (c) のように多少周期性があってもそうでないと判断してしまうことがある。しかしこの判定はそれほど厳密なものである必要はない。なぜなら残差信号に周期性があるのに、誤って原音声信号に対して音源情報抽出を行っても、それに続く振幅包絡特性からの音源情報抽出アルゴリズムが本質的に原音声信号及び残差信号のどちらに対しても有効なアルゴリズムであるために、音源情報を抽出できないということはないためである。ただし残差信号からの方がより正確な音源情報を抽出できるという意味で、この判定結果が有効になってくる。そして実験結果によってこの方法が十分に簡単で有効な方法であるということがわかった。

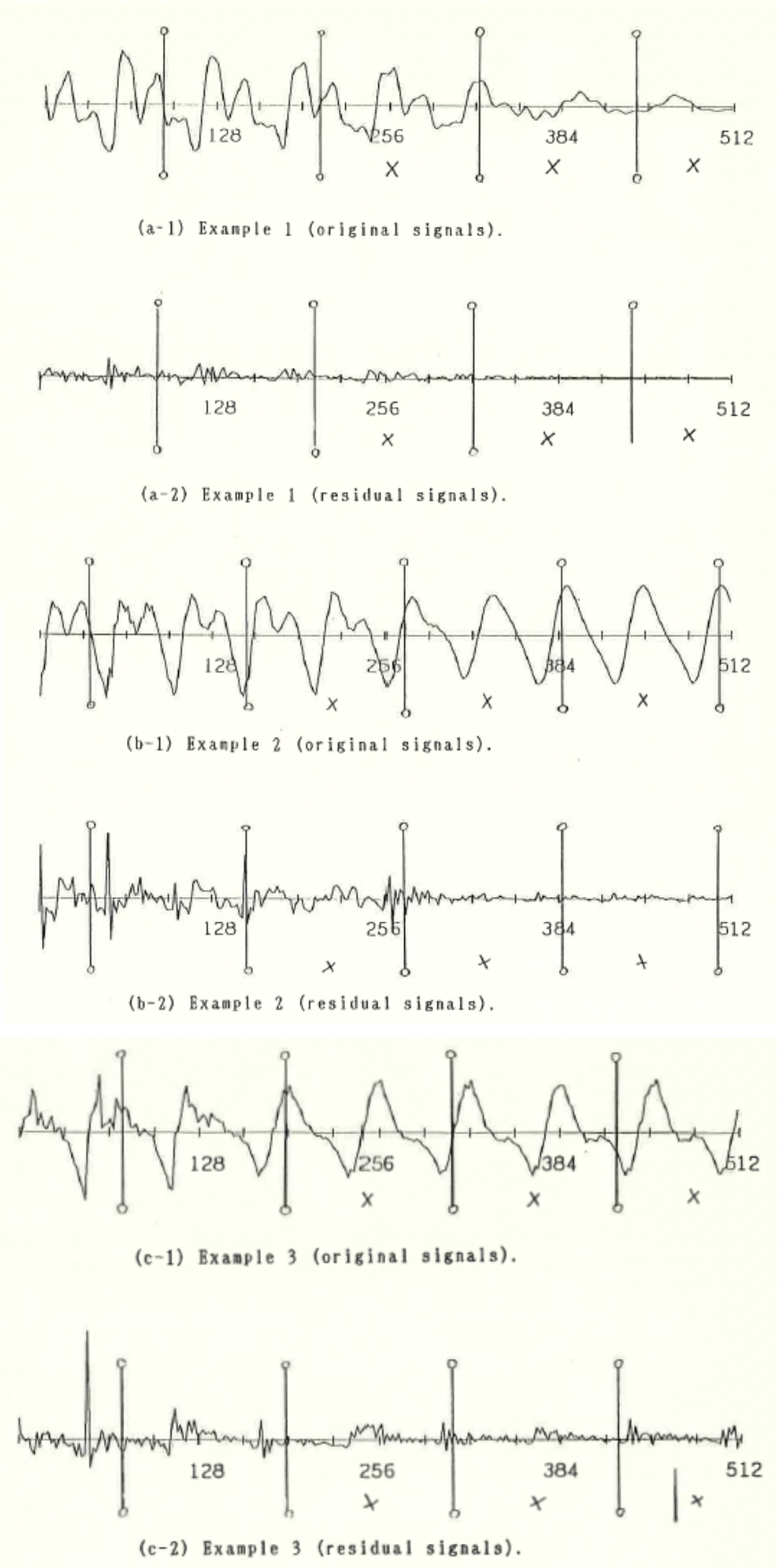


図 2.11 残差信号と原音声信号の判定

Fig. 2.11 Decision of residual signals or original signals.

2.7.3 振幅包絡特性

次に上記判定の後、2.5 節で述べた図 2.1 に示したアルゴリズム、及び図 2.9、図 2.10 に基いて残差信号又は原音声信号に対して計算される振幅包絡特性について、その結果を示す。まず、2.6 節の説明で用いた図 2.8(c) は音声ファイル No.9(XVOI02.DAT) の残差信号の振幅包絡特性の例である。各ピッチピークの位置と振幅を非常に正確にモデル化していることがわかる。同図 (b) は音声ファイル No.3(DATA18.DAT) の残差信号の振幅包絡特性の例である。これは「クリスマスは・・・」の「ク」の部分に当たり、無声音から有声音へのわたり部分であるが、ピッチピークの位置と振幅をやはり良く近似している。また同図 (d) は音声ファイル No.5(DATA28.DAT) の残差信号の例である。この場合はピッチピークの他に疑似ピークが現れている。更に同図 (a) は音声ファイル No.3 の残差信号の他の部分の例である。これは「クリスマスは・・・」の「ク」の始めの無声音部分であり、意味のない疑似ピークが現れていることがわかる。

次に図 2.12(a) は、音声ファイル No.9(XVOI02.DAT) の原音声信号の振幅包絡特性の例である。この場合も各ピッチピーク位置に正確に対応しており、この振幅包絡特性を求めるアルゴリズムが、原音声信号 (正弦波状の場合) にそのまま適用できることがわかる。なおピークの振幅は多少ずれているが、原音声信号の場合はピークの振幅はその位置の近隣の残差信号の平均振幅を用いるので問題はない。同図 (b) は、音声ファイル No.5 の他の部分の原音声信号に対するものである。この場合、接近して疑似ピークが現れている。また同図 (c) は、音声ファイル No.9 の他の部分の原音声信号に対するものである。これは「・・・ますか」の「ま」の始まり部分であり、疑似ピークはあるがピッチピークを良く近似していることがわかる。以上のようにして求まる振幅包絡特性は、残差信号に対する場合、原音声信号に対する場合の両方において、疑似ピークの出現を最小限に抑え、かつピッチピークを非常に正確に近似することがわかる。なお図 2.1 の各図は、負の振幅部分から求めたが、これは後続の母音部分の残差信号のピークの特性と合わせたためである。

時間域音声信号の振幅包絡特性を全極モデルで近似するという考えは、本手法における最大の特徴である。図 2.6 を見ればわかるとうり、ピッチパルス以外の部分の影響が非常に有効に除去されており、しかもピッチパルス部分は非常に正確に近似されていることがわかる。これは全極モデルで振幅包絡特性を求めているために、ピッチパルスによる鋭いピークも有効に近似されているからである。しかし、ここで重要なことは全極モデルの次数をどれくらいに設定したらよいかということである。本手法では 1 分析区間の時間域残差信号を直流分から最高周波数までの疑似スペクトルとみなしており、一方、2.6 節で述べたように全極モデルの次数はスペクトルのピークの数の 2 倍が最適次数であるため、1 分析区間内のピッチピークの数の 2 倍が最適次数となる。しかしピッチピークの数は変化するため、最大ピーク数を近似できるだけの次数とした。そのために図 2.8(d) に示すように、ピッチピーク以外の部分にもピークが現れてしまうのである。しかし、疑似ピークは真のピークの間に 1 個程度しか現れないということが実験的にわかっているので (これは次数に制約があるということからもわかる)、それに続く訂正アルゴリズムが非常に簡単なもので済むのである。この場合、次数が低すぎるとピッチピーク以外の影響を受け、ピーク部分が正確に近似されなくなるので、あえて次数を高くして疑似ピークが出ることを許したのである。また疑似ピークはホルマントによる影響を近似する場合が多いので、真のピークの直後に偏って出現する場合が多く、その性質からも訂正アルゴリズムを構成しやすくしているといえる。

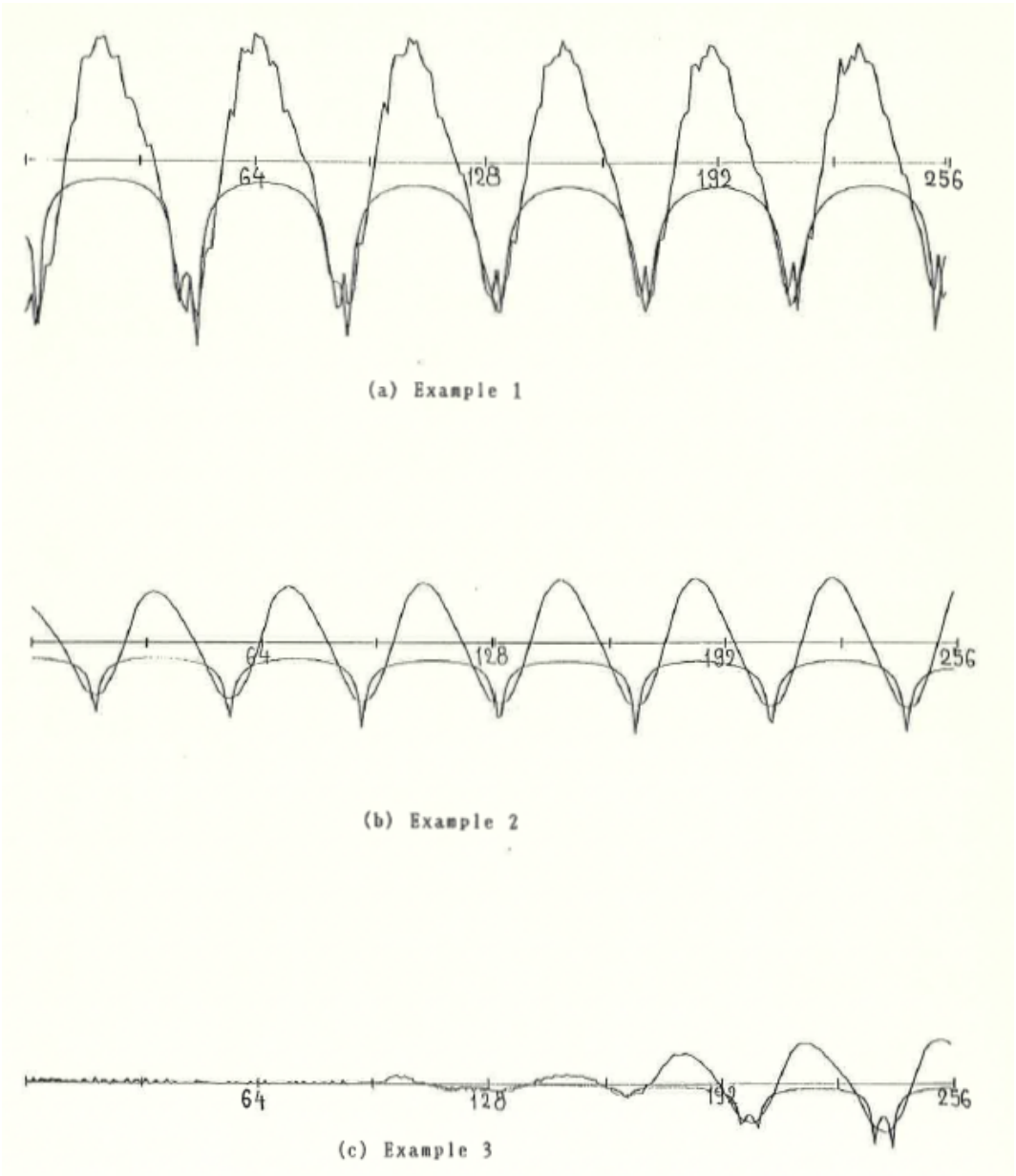


図 2.12 残差信号の振幅包絡特性

Fig. 2.12 Amplitude envelope characteristic of original signals.

次に上記の考えは原音声信号に対しても同様に適用できた。この場合の極によるピークは図 2.12 に示すように波形のピーク部分にうまく現れ、ピッチのピークにして良いといえる。さらにこの場合において分析されるべき原音声信号は、前段の判定アルゴリズムによって、正弦波に近い信号に限定されているため、ピッチ以外の影響があまりなく、それによって振幅包絡特性からのピーク抽出がうまくできるといえる。ただしこの場合は、原音声信号を半波整流してから行うというのが重要な点である。

2.7.4 ピーク訂正とピーク抽出精度

上記振幅包絡特性の計算結果に基いて、2.6 節のアルゴリズムに従ってピーク抽出・訂正を行い、音源情報を抽出した結果について示す。

まず予備実験によって、ピーク訂正のための各閾値を次のように決定した。始めに振幅判別ルーチン I − (i) の値 TH_{cut} は、残差信号の場合、$TH_{cut} = 60$、原音声信号の場合、$TH_{cut} = 150$ と

した。次に振幅判別ルーチンI − (ii) の値 TH_r は、残差信号の場合、$TH_r = 2.0$、原音声信号の場合、$TH_r = 4.0$ とした。続いて間隔判別ルーチンの平均値に対する許容誤差範囲は±15%とした。

上記パラメータを用いて、図 2.1 の音源情報抽出システムに従って、前記音声 10 ファイルについて音源情報を抽出した。そしてその結果について、グラフィックディスプレイを用いた原音声波形及び残差波形からの視察による抽出結果を基準にして評価を行った。その結果を表 2.1 に示す。

表 2.1 ピーク抽出精度

Table 2.1 Accuracy of peak extraction.

ファイル番号	抽出精度 (%)
1	98.6
2	97.2
3	97.5
4	97.1
5	99.2
6	94.7
7	98.3
8	96.5
9	98.1
10	100.0

ここで抽出精度とは、抽出されたピークの位置が、視察による抽出位置に対して ±2 サンプル以内の誤差で一致する割合を、パーセントで示したものである。この結果より、各音声ファイルとも非常に高い抽出精度であった。ただし、音声ファイル No.6 のみ、多少精度が悪かった。また No.10 については 100%の抽出精度であった。

次に誤りが起こった原因について各原因別についてまとめたのが表 2.2 である。

表 2.2 ピーク抽出誤り原因

Table 2.2 Cause of peak extraction errors.

(単位：件数)

誤り原因	欠落		誤抽出	
分析対象	原波形	残差波形	原波形	残差波形
回数	29	24	0	5

これより、振幅包絡にピークが現れず、その部分のピッチピークが欠落してしまったものが多かった。また、現在のピークの次にピークがなく (語尾)、そのピークがホルマントによる疑似ピークであり、間隔判別ルーチンで判別できなくても、それに続く振幅判別ルーチンで誤判別され抽出されてしまう場合があった。なお前記欠落によるものは、正弦波状音声部分に近いものが多かった。

ピーク振幅の抽出精度については図 2.8 などから分かるように、全ての場合について残差信号のピークの実際の振幅と比べて、10%程度の誤差しかなかった。

ピーク訂正は振幅包絡特性の性質をうまく使ったものである。まず間隔判定ルーチンは、ピッチ間隔 (周期) が連続的に変化するということを利用したものである。即ち、実験では隣合うピッ

チ間隔が 15%以上変化するということは有り得ないと仮定しており、これは実験の結果正しいということが確かめられた。

次にこれだけでは音声の始まりや、何らかの原因で誤ったピッチ間隔が抽出された場合正しいピーク抽出をすることができないので、そのための処理として振幅判定ルーチンを用いるのである。そしてこのルーチンは、図 2.8 などから分かるように真のピーク部分はその直後の谷部分と振幅差が大きくなるという振幅包絡特性の特徴をうまく利用したものである。更に、疑似ピークは前記したように真のピークに接近して現れるため、直前の谷部分との振幅差が小さくなり、これによってうまく判別を行えるのである。またこのことは、図 2.12 などの原音声信号の場合にも同様に考えられるといえる。なお、振幅判別ルーチンⅠ－ (i) の値 TH_{cut} は、残差信号の場合 60、原音声の場合 150 と設定したが、これは、これらの値以下の振幅部分は有効な音声ではないとしているためである。この場合入力音声は、AGF(自動利得調整器) などによって、最大値が 12bit 量子化精度の最大値 2048 付近になるように調整されているという仮定のもとで上記☒閾値を設定してある。

次にピッチ抽出精度について見ると、表 2.1 から見てわかるように高い抽出精度を得ることができた。この精度は全ピーク数に対する正解率であるので、非常に精度が良いといえる。そして、全極モデルによるピッチピーク部分のモデル化が、非常に有効なものであるということを示しているといえる。また表 2.2 より、誤りは原音声が正弦波状の音声の場合に多い。即ち、2.7.2 の判定結果で残差信号にあまり周期性がないのに、周期性があると判定され、残差信号に対して音源抽出が行われたために、振幅包絡特性にその部分のピークが現れなかったためであると考えられる。また、原音声に対して音源抽出が行われた場合でも、振幅包絡特性にピークが現れない場合があり、これは、原音声信号中にピッチピークの他にホルマントによるピークが現れ、ピーク数に対する分析次数が相対的に低くなってしまい、ピーク部分が正しくモデル化できなかったためであると考えられる。また残差信号から音源抽出を行う場合で、語尾のピークを抽出する際に、振幅判別ルーチンが働いてしまい、疑似ピークを真のピークとして抽出してしまう場合があるようである。

しかし上記のような誤りは非常に回数が少なく、本システムは全体的にみて優れた性能を持つものであるといえる。

2.7.5 合成による音質評価

次に、音源情報抽出システムを用いて、実際に音声合成を行った結果について示す。まず、音源情報抽出システムの各条件は、2.7.1 と全く同様であり、このシステムは、有声音源情報 $V(P_j、A_{mj})$、又は無声音源情報 $UV(A_{me})$ を出力する。

次に音声合成系は PARCOR 合成系であり、次数は 10 次、分析フレーム長は 256 サンプル、PARCOR 係数は 16 サンプル間隔で直線補間を行った。

そして音源情報が有声音源情報の時は、256 サンプル毎について、ピッチ位置 P_j に振幅 A_{m_j} のインパルスを立てて残差信号を合成し、PARCOR 合成フィルタへの入力音源とした。また無声音源情報の時は、振幅 A_{me} で 256 サンプルのホワイトノイズを生成し、入力音源とした。

以上の条件で対比方式をグラフィックディスプレイを用いた、視察による 1 分析フレーム (256 サンプル) 毎の平均ピッチ周期抽出法とし、対比較受聴試験を行い、プレファレンススコアをとった。また入力音声ファイルは、2.7.1 で説明した No.8、及び No.9 である。試聴試験者は 6 人で、各ファイルとも対比方式による合成音声と組にして 3 回づつランダムに発声させた。

その結果、本方式による合成音声の方がよいと答えた比率が 61.1%であり、本方式の有効性が

確かめられた。

また試聴者に意見を聞いた結果、次のような評価を得た。

(i)対比方式の方がこもっている。本方式の方がはっきりしている。
(ii)ピーク抽出エラーによるものと思われるクリックノイズが、本方式の方で少し聞こえる。
(iii)対比方式の方が、合成音にメリハリがない。

以上の結果より、本手法の目的の 1 つである合成音声の品質向上という点について、有効な結果が得られたといえる。試聴者の評価では、本方式の方がはっきりしているという意見が多かった。これは本方式が、残差信号中のピッチピークの位置と振幅を正確にモデル化していることに帰因していると思われる。即ち、残差信号におけるピッチインパルス列は、完全に周期的なものではなく、1 分析フレーム内 (256 サンプル) でも微妙に変化しており、その変化が逆特性フィルタによって除去しきれなかった周波数特性となって残っている。従来方式では、そのような微妙な変化をモデル化することはできず、それによる本方式との違いが合成音声の明瞭性に現れたと考えることができる。また本方式は、音源情報を抽出しにくい鼻音部分などについても正確に求まっているため、合成音声の原音声に対する再現性が高いということも、良品質の合成音であるという結果に好影響を及ぼしているといえる。

更に、「スキーは南国でできますか」の「キ」や「か」の部分がはっきりとしているという意見もあり、これは声音部分から有声音部へのわたりにおける音源情報のモデル化がうまく行われ、それによって合成音声の音質が良くなったと考えられる。

以上の結果より本システムの音声合成への適用は、非常に有効なものであるということが確かめられた。ただし抽出誤りによるクリックノイズなどは、現段階では完全に自動的には除去できなかった。

2.7.6 分析フレーム長と乗算回数

図 2.1 において、振幅包絡信号を得るために必要な乗算回数は、分析フレーム長を N サンプル、LPC 分析の次数を P 次とすると、各処理で次のように与えられる [37][39]。

2 乗 (2 回) :N ×2
逆 FFT :$\frac{1}{2} \times \frac{2N}{4} \times \log_2 \frac{2N}{4} \times 4$
(実数部のみで偶対称データなので $\frac{1}{4}$ サイズの FFT で良い。)
LPC 分析 : $P^2 + 3P$
FFT : $\frac{1}{2} \times \frac{2N}{2} \times \log_2 \frac{2N}{2} \times 4$
2 乗和 (図 2.10(a)) :2N

以上の計算をもとに、256 サンプルからピーク抽出のための振幅包絡信号を計算するために必要な分析フレーム長と分析次数及び計算回数の関係は、表 2.3 のようになる。

表 2.3 分析フレーム長、次数、及び乗算回数の関係
Table 2.3 Relations between frame length,analysis order,and numbers of multiplications.

分析フレーム長	LPC 次数	乗算回数
256	24	7560
128	12	6504
64	6	5592
32	3	4752

これによると分析フレーム長が短くなると、乗算回数が減ることがわかる。しかし分析フレーム長が短くなると、ピッチピークが存在しないフレームが多くなり、疑似ピークが多く抽出され、抽出精度が悪くなってしまう。簡単な検討の結果、1 分析フレームを 128 サンプル及び 256 サンプルとした場合に、ほとんど同精度でピークが抽出された。従って、1 分析フレームは 128 サンプルが望ましいといえる。

以上の計算を含めた本システムによる処理量は、従来方式と同等又はそれ以下であり、FFT や LPC 分析アルゴリズムがハード化されつつある現在、音声分析合成系への実装は、十分可能であるといえる。

2.8 総括

第 2 章では、マクロな音源情報の新しいモデル化を提案した。

そのモデル化に基づく音源情報抽出法として、時間域信号を周波数スペクトルとみなし、その振幅包絡特性を LPC に基く全極モデルとして求める方法を提案した。

更に、鼻音部分を識別し、その部分の処理を適切に行う方法を提案した。

これらの方法により、ピッチピークの位置と振幅を正確に求めることが可能となり、このシステムを音声合成系に適用した結果、無声音と有声音のわたり部分の情報やピッチ周期の微妙な変化を正確にモデル化でき、合成音声の音質が向上することを明らかにした。

第 3 章　LPC 有声音残差のピッチ同期分析に基く零点を有する励振パルスモデル

3.1　概要

第 2 章で提案した手法により、LPC 音声分析合成システムのためのマクロな音源情報の正確な抽出を実現した。これにより、1.2kbits/sec〜4.8kbits/sec の低ビットレートの伝送帯域において十分に高能率な音声符号化の達成が可能であることを示唆した。

一方において第 1 章では、4.8kbits/sec〜9.6kbits/sec の低・中ビットレートの音声符号化のために LPC 方式を利用しようとした場合、残差信号、特に有声音残差から、マクロな音源情報の抽出に加え、ミクロな音源情報を抽出することが必要であることを明らかにした。また、残差に含まれるミクロな音源情報は、LPC 方式における全極モデルとの差としての零点特性であることを述べ、そのモデル化が緊急に必要な課題であることを明らかにした。

本章では、ミクロな音源情報として有声音残差から上記零点特性を抽出するための分析手法として、LPC 有声音残差のピッチ同期分析が有効かつ実現可能であり、それによるピッチ同期振幅スペクトルが極めて特徴的な零点を有することを示す。

そしてこの分析に基く有声音源モデルとして、有声音残差のピッチ同期振幅スペクトルを逆 LPC 分析によってモデル化し、その係数から上記スペクトルを再合成することにより得られる零点を有する励振パルス (ZEP) を提案する。

これをもとにして有声音残差のモデル化を実際に行い、それより再合成された残差を有声音源として LPC 合成して得た合成音声の主観的評価を行うことにより、本手法の有効性を確認する [40]。

3.2　有声音残差のピッチ同期分析

3.2.1　有声音残差のピッチ同期離散的フーリエ変換

1.4 節での議論より、有声音残差 $\epsilon(n)$ は LPC 分析において、全極フィルタとの差としての零点特性を有する準周期的なパルス信号列であると考えられる。そこで、有声音残差 $\epsilon(n)$ を次の 3.1 式により定義する。

$$\epsilon(n) = \sum_{r=-\infty}^{\infty} h_e(n+rN_p) \qquad (-\infty < r < +\infty) \tag{3.1}$$

ここで h_e は前記特性を有するインパルス応答であり、N_p はピッチ周期である。また、h_e のフーリエ変換を $X(\mathrm{e}^{j\omega})f$ を、次の 3.2 式により定義する。

$$X(\mathrm{e}^{j\omega}) = \sum_{m=-\infty}^{\infty} h_e(m)e^{-j\omega m} \tag{3.2}$$

今、$0 \leqq n \leqq N_p - 1$ の範囲で方形窓を $\epsilon(n)$ に乗じて 1 ピッチ周期を切り出した範囲で、$\epsilon(n)$ のフーリエ変換 $\tilde{X}(k)$ は、次の 3.3 式により計算される。

$$\tilde{X}(k) = \sum_{n=0}^{N_p-1} \epsilon(n) \mathrm{e}^{-j\frac{2\pi}{N_p}kn} \tag{3.3}$$

3.1 式を 3.3 式に代入すると、次の 3.4 式となる。

$$\begin{aligned} \tilde{X}(k) &= \sum_{n=0}^{N_p-1} [\sum_{r=-\infty}^{\infty} h_e(n+rN_p)] \mathrm{e}^{-j\frac{2\pi}{N_p}kn} \\ &= \sum_{r=-\infty}^{\infty} [\sum_{n=0}^{N_p-1} h_e(n+rN_p)] \mathrm{e}^{-j\frac{2\pi}{N_p}k(n+rN_p)} \end{aligned} \tag{3.4}$$

ここで、$m = n + rN_p (-\infty < m < +\infty)$ とおいて 3.4 式を書き換えると、次の 3.5 式と書ける。

$$\tilde{X}(k) = \sum_{m=-\infty}^{\infty} h_e(m) \mathrm{e}^{-j\frac{2\pi}{N_p}km} \tag{3.5}$$

3.2 式と 3.5 式より、次の 3.6 式が得られる。

$$\tilde{X}(k) = X(\mathrm{e}^{j\frac{2\pi}{N_p}k}) \tag{3.6}$$

これより $\epsilon(n)$ のピッチ同期離散的フーリエ変換は、$X(\mathrm{e}^{j\omega})$ を $\omega = \frac{2\pi k}{N_p}$ でサンプリングしたものに等価となる。即ち上記各サンプル点における成分は、インパルス応答 h_e の周波数スペクトルを正確に表しており、ピッチの調波構造の影響を受けずに、その零点特性の正確な観測が可能となる。

3.2.2 有声音残差の各ピッチ周期の開始点の決定

有声音区間の原音声信号のピッチ同期分析を行う場合、各ピッチ周期を切り出すためにその開始点を決定する必要がある。しかし一般に原音声の場合、図 3.1 に示すようにその開始点を決定するのは困難であり、また開始点の振幅を小さくすることも困難である。そのために従来は、ピッチ同期分析の分析・合成システムへの適用は難しかった [41][42]。

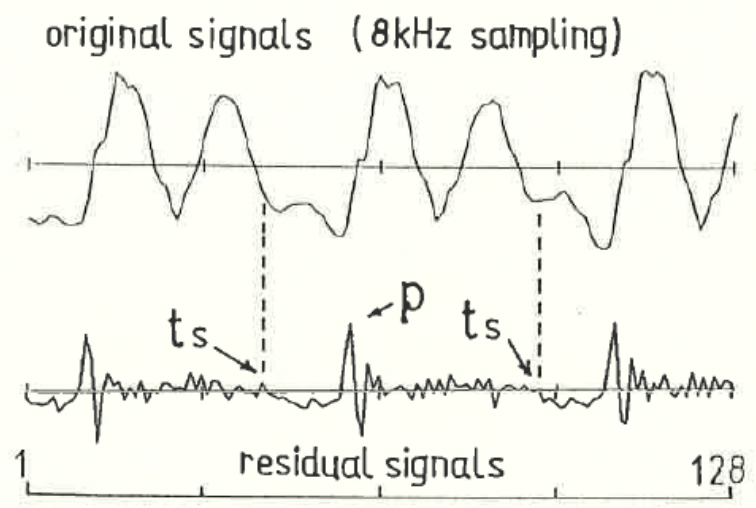


図 3.1 有声音原音声信号と残差信号

Fig. 3.1 Voiced original and residual signals.

これに対して、有声音残差は全極モデルによる調音特性が除かれているため周期性が明確に現れ、そのピーク p を自動抽出するのは第 2 章で述べた手法等により可能であり、また各ピッチ開始点も視察等により比較的容易に抽出しやすい。そして、各ピーク位置とピッチ開始点 ts の関係は、3.2.4 で後述するように音声の閉鎖速度に依存し統計的な傾向を有するため、その関係を用いて図 3.1 に示す有声音残差から、各ピッチ周期の開始点を決定する事ができると考えられる。これに加えて、開始点付近の残差パワーは、十分に減衰しているため、方形窓を用いるピッチ同期分析の場合、フレーム端の影響を最小限に抑制でき、切り出し位置の誤差に対する感度が低いという利点を有する。

この事実により、有声音残差のピッチ同期分析は、3.2.1 で述べた効果の他に、残差分析の自動化を可能にし、これを用いた分析合成システムの実現の可能性を示している。

本手法においては、まず各ピッチ周期を視察により切り出して有声音残差のピッチ同期振幅スペクトルを詳細に観測し、その後各ピッチ区間におけるピーク位置とピッチ周期の統計的分析から、各ピッチ開始点を決定しその影響を調べる。

3.2.3 有声音残差のピッチ同期スペクトル

まず、有声音残差の各ピッチ周期を 3.2.2 で述べたように視察により切り出し、3.2.1 で述べた分析法に基いて有声音残差のピッチ同期振幅スペクトルを求める。ここで、3.2.1 で述べた分析法によると、周波数軸上のサンプリング位置は 3.6 式に示すようにピッチ周期 N_p に依存する。しかし、ピッチ同期スペクトルを統計的に分析し、又はモデル化することを考えると、周波数軸上のサンプリング間隔が一定である方が望ましい。そこで本手法では、切り出した各ピッチ周期の N_p サンプルを 256 点 FFT の実数部の始めの N_p 点に挿入し、残りの実数部及び虚数部には 0 を挿入して分析を行う。この分析法により、周波数軸上のサンプリング間隔を△ $\omega = \frac{2\pi}{256}$ の等間隔にすることができる。この時得られるピッチ同期振幅スペクトルは、厳密には 3.2 式の $X(\mathrm{e}^{j\omega})$ を $\omega = \frac{2\pi}{256}$ でサンプリングしたものとは異なるが、近似的な補間処理を実質的に少ない誤差で等価的に行うことができ、また計算時間の面からも FFT を用いているため有利な手法であり、ピッチ同期分析の自動化を可能にするものである。

図 3.2 は、上記分析法による男性及び女性に対する 5 母音の残差の典型的なピッチ同期振幅スペクトルである。このスペクトルは、図 1.2 に示したような従来の短時間スペクトルとは大きく異なり、非常に特徴的な零点を有することが分かる。そして、零点の帯域幅は全極モデルによるモデル化の影響を受け音韻により変化するが、その周波数は音韻によってあまり大きな影響は受けず、話者により異なる傾向がある。これは、1.4 節において述べたように有声音残差の零点特性が声帯及び鼻音の調音機構に起因し、全極モデルとの差として与えられることからも予想される結果である。なお周波数的な特徴については、3.3 節の有声音残差のモデル化において更に検討を加える。

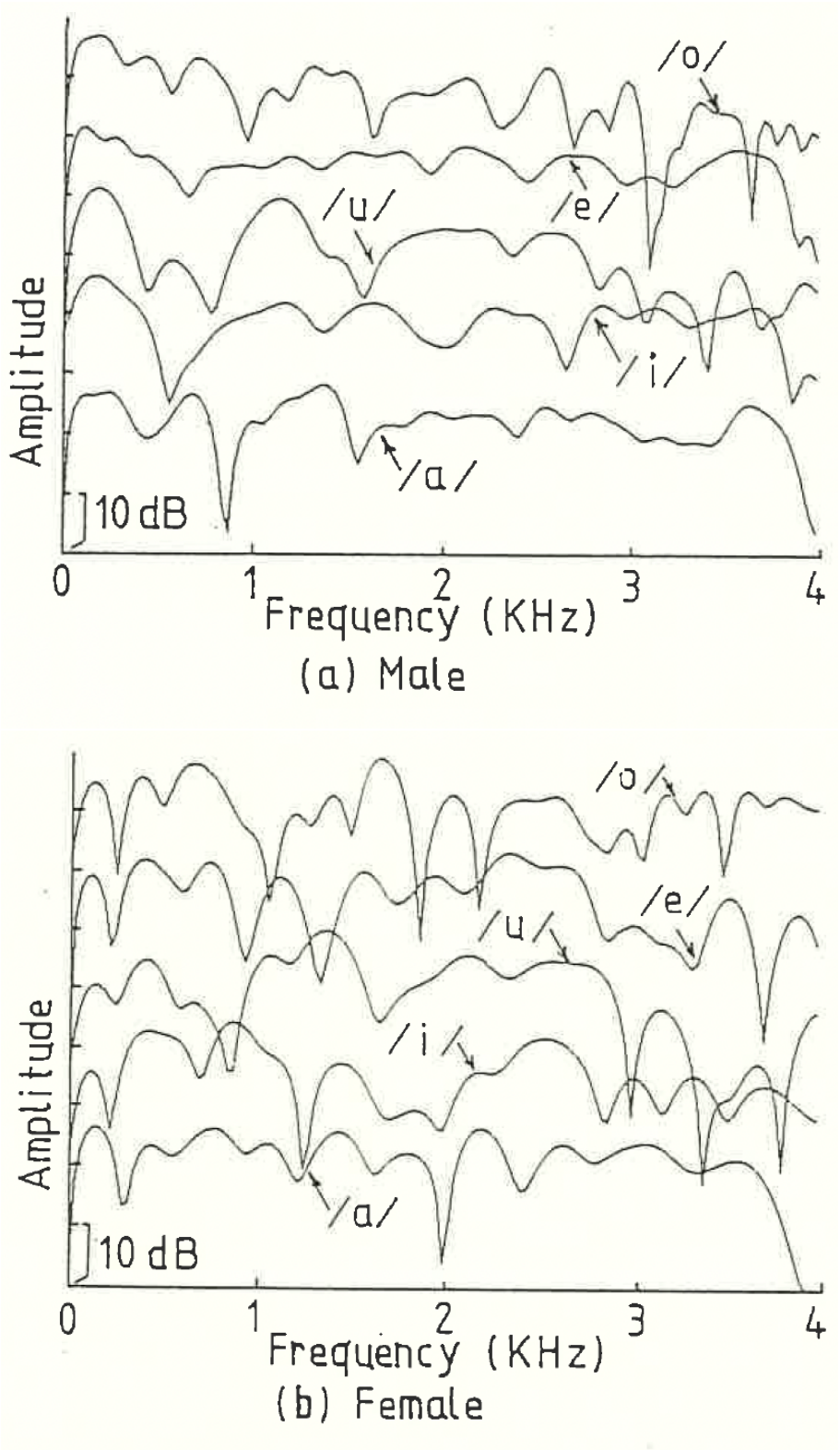


図 3.2 FFT 計算による 5 母音の残差のピッチ同期振幅スペクトル

Fig. 3.2 FFT-computed pitch synchronous amplitude spectrum of residual signals of five vowels.

次に、図 3.3 は有声音「あ」の残差の連続する 5 ピッチ周期のピッチ同期振幅スペクトルである。同図より連続するピッチ周期間で、特に低域から中域の周波数帯において Q の大きな準定常的な零点が現れている。

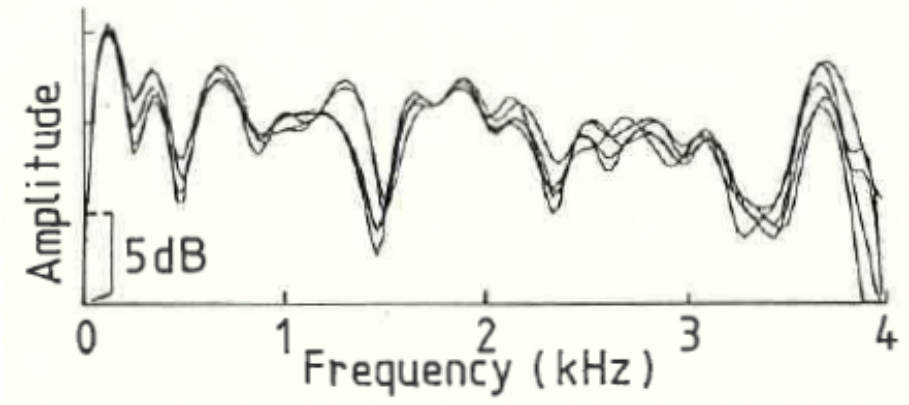


図 3.3 連続ピッチ区間の残差のピッチ同期振幅スペクトル

Fig. 3.3 Pitch synchronous amplitude spectrum of residual of continuous pitch periods.

次に上記定常性の定量的な分析を行う。まず、有声音残差の連続する J ピッチ周期からなる区間をフレーム m とし、フレーム m 内の各ピッチ周期 $j(1 \leqq j \leqq J)$ の 256 点 FFT によるピッチ同期対数振幅スペクトルを $Z_k(m,j)$ dB(k：周波数、$1 \leqq k \leqq 129$) とし、フレーム m における $Z_k(m,j)$ の平均スペクトル $\mu_k(m)$ dB を次の 3.7 式で定義する。

$$\mu_k(m) = \frac{1}{J}\sum_{j=1}^{J} Z_k(m,j) \qquad \text{dB} \tag{3.7}$$

そしてフレーム m において、各ピッチ周期毎の上記 $\mu_k(m)$ とのスペクトル歪 (2 乗偏差)$Z(m,j)$ dB^2、及びその標準偏差 $\sigma(m)$ dB を夫々、次の 3.8 式及び 3.9 式で定義する。

$$Z(m,j) = \frac{1}{128}\sum_{k=1}^{128}\{Z_k(m,j) - \mu_k(m)\}^2 \qquad \text{dB}^2 \tag{3.8}$$

$$\sigma(m) = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J} Z(m,j)} \qquad \text{dB} \tag{3.9}$$

更にこの標準偏差 $\sigma(m)$ を、1 ピッチ周期ずつ更新して得た全フレーム M について平均し、平均標準偏差 $\bar{\sigma}$ dB を次の 3.10 式で定義する。

$$\bar{\sigma} = \frac{1}{M}\sum_{m=1}^{M} \sigma(m) \qquad \text{dB} \tag{3.10}$$

上記平均標準偏差 $\bar{\sigma}$ は連続する J ピッチ周期でピッチ同期振幅スペクトルがどの程度歪むかを、スペクトル歪の標準偏差の形で表したものであり、ピッチ同期振幅スペクトルの定常性を表していると考えられる。

表 3.1 に、男女各 1 名の連続音声「あいうえお」の残差のピッチ同期振幅スペクトルについて、J=5 とした場合について上記 $\bar{\sigma}$ を求めた結果を示す。

表 3.1 残差のピッチ同期振幅スペクトルの定常性

Table 3.1 Constancy of pitch synchronous spectrum of residual.

性別	平均標準偏差 σ
男	2.31　dB
女	3.15　dB

これより連続する 5 ピッチ程度 (30msec 前後) では、ピッチ同期振幅スペクトルはあまり変化せず、準定常的であることがわかる。この結果は、ピッチ同期振幅スペクトルのモデル化を行った場合に、モデルのパラメータが安定に抽出できることを示し、将来的に 5〜30msec の間でパラメータの補間が可能であると考えられる。

図 3.2、3.3、及び表 3.1 の結果により、有声音残差のピッチ同期振幅スペクトルを、特に零点に重みをおいてモデル化することにより、合成音声の音質を高めることができると考えられる。

3.2.4　各ピッチ周期のピークと開始点の関係

一般に有声音源の声帯の振動波形 (声門パルス) は、図 3.4 に示すようなモデルで表すことができる [42]。

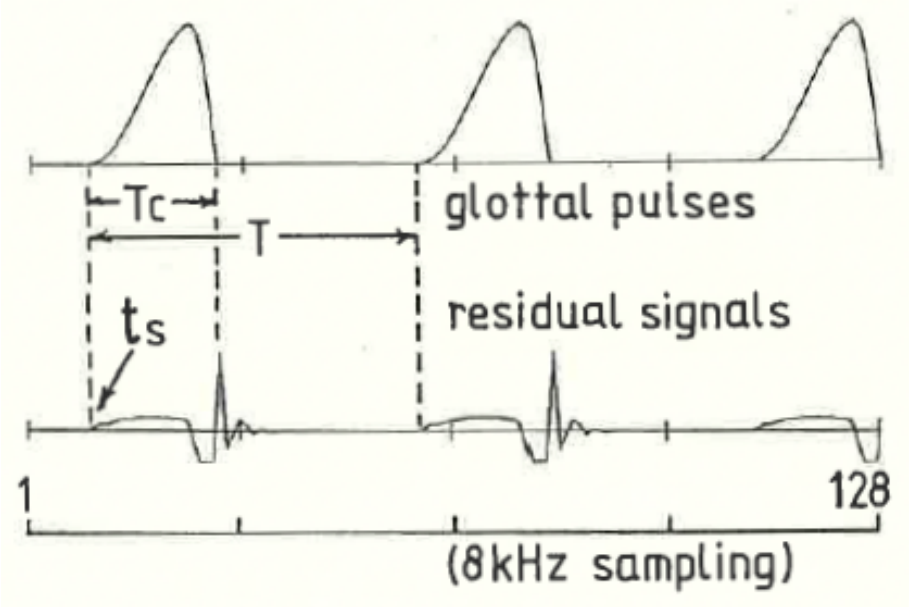


図 3.4 声門パルスと残差信号の関係

Fig. 3.4 Relation between glottal pulses and residual signals.

この場合の声門パルスの全体的な周波数特性 (包絡特性) は、LPC 分析においては声道特性と共に LPC 係数に吸収される。そしてこの声門パルス波形から、その周波数特性を LPC 逆フィルタにより平坦化して得た残差波形は、同図に示すように LPC 分析における一般的な有声音残差波形に良く一致する。そしてこの時の鋭いピークは声門パルスとの位相関係から、声帯が急激に閉鎖する瞬間に現れることがわかる。

今、図 3.4 において、1 ピッチ周期 T と声帯が開いている区間 Tc との比である duty factor (Tc/T) は、ピッチ周期 T と発声のストレスに関連して統計的な傾向を有していることが知られている [43]。従って、有声音残差におけるピークが同図に示すように声門パルス波形と密接な関係にあることより、有声音残差においても上記 duty factor は統計的な傾向を有すると考えられ、そ

の傾向がわかれば、ピーク位置からピッチ開始点 t_s を近似的に決定することができる。表 3.2 は 3.2.3 節において用いた男女各 1 名の連続音声「あいうえお」(各約 320 ピッチ周期) において、視察によって求めた各ピーク位置とピッチ開始点から、上記ピーク位置対ピッチ周期の比の平均と標準偏差を計算した結果である。

表 3.2 有声音残差の duty factor

Table 3.2 The duty factor of voiced residual.

性別	平均	標準偏差
男	42.29 %	0.87 %
女	42.53 %	0.85 %

表 3.2 より上記音声の場合、ピーク位置はピッチ開始点からピッチ周期のほぼ 42%の付近に現れることがわかった。次に、上記 duty factor を用いて各ピーク位置からピッチ開始点を決定し、ピッチ同期分析を行った場合の振幅スペクトルへの影響を調べてみる。そのために視察により求めたピッチ開始点を用いて、3.2.3 と同様に 256 点 FFT により計算したピッチ同期対数振幅スペクトル PZ_k と、各ピーク位置から表 3.2 の値を用いてピッチ開始点を決定し、同様にして計算したピッチ同期対数振幅スペクトル PZ_k とを用いてスペクトル歪 SD を次の 3.11 式で計算し、それを全ピッチ周期において平均することにより平均スペクトル歪 $\bar{SD}$ を求めた。

$$\bar{SD} = \frac{1}{128}\sum_{k=1}^{128}(PZ_k - PZ'_k)^2 \tag{3.11}$$

表 3.3 は表 3.2 の計算時と同様の音声を用いて $\bar{SD}$ を計算した結果である。

表 3.3 duty factor 固定により生ずる平均スペクトル歪

Table 3.3 Mean spectral distortion when duty factor was fixed.

性別	平均スペクトル歪 $\bar{SD}$
男	2.42 dB
女	2.43 dB

これより各ピッチ周期の切り出し誤差によるピッチ同期振幅スペクトルへの影響は、人間の聴覚特性が零点に比較的鈍感であることを考えると、十分に小さい範囲であることがわかる。

これは 3.2.3 で述べたようにピッチ開始点付近の残差パワーが十分に減衰しているためであると考えることができる。

3.3 有声音残差のモデル化 - Zero Exitation Pulse Model

次に、3.2.3 及び 3.2.4 の分析結果に基いて、有声音残差のピッチ同期振幅スペクトルのモデル化を行う。このスペクトルは図 3.2 などからわかるように非常に特徴的な零点を有し、かつそのスペクトルは滑らかであり、零点に重みをおいたモデルとして様々なものを考えることができるが、ここでは逆線形予測モデルを用いる。これは入力スペクトルが全零型の場合に、そのスペクトルを反転して全極型とした後 LPC 分析を行い、それから求まる LPC 係数を入力スペクトルのモデルパラメータとするものである。合成時には全く逆の操作を行うことによりスペクトルを再生できる。

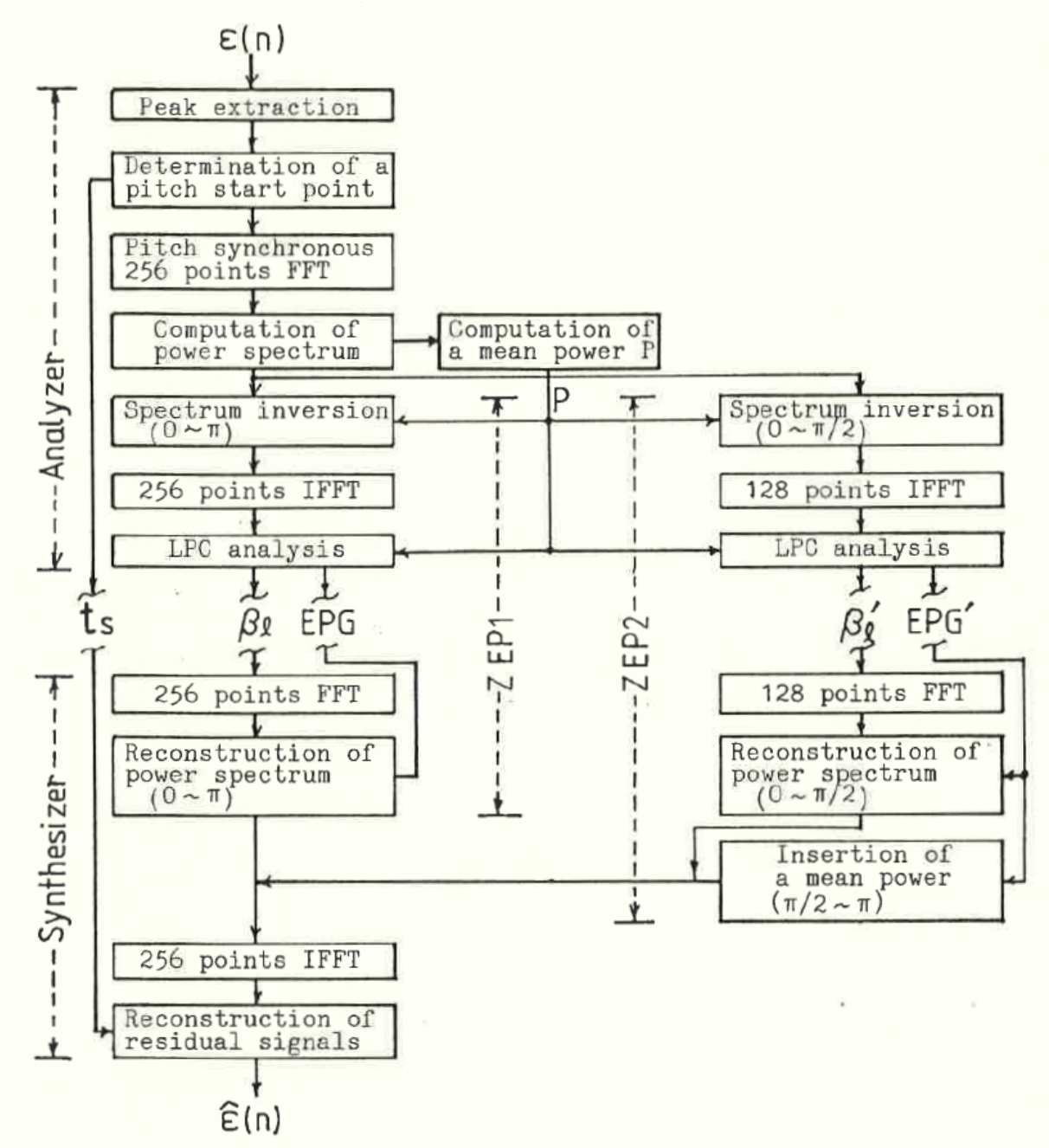


図 3.5 ZEP を用いた分析・合成システムのブロック図

Fig. 3.5 Block diagram of the analysis synthesis system using ZEP.

しかしこの方法においては前処理として、入力音声からまず極の影響を取り除いた後、スペクトルの平滑化を行ってピッチの調波構造の影響を除去する必要があるが [15]、1.5 節で述べたように従来の一般的な短時間フーリエスペクトルにおいては元々零点特性がぼかされてしまっているため、零点特性に重みをおいたスペクトル平滑化は困難であった。

これに対して有声音残差のピッチ同期振幅スペクトルは、LPC 逆フィルタにより極の影響が除かれており、更にピッチ同期分析を行うことによりピッチの調波構造の影響を除去しているため零点特性が明確に現れ、従って上記逆線形予測モデルを容易に適用できる。

以下、モデル化の具体的手順について述べる。図 3.5 はモデル化のためのブロック図である。ここでは 2 つのモデルを提案するが、まずモデル 1 について説明する。始めに有声音残差から各ピッチ周期のピークを抽出した後、表 3.2 で求まった duty factor を用いてピークの開始点を決定し 3.2.3 と同様に 256 点 FFT によりピッチ同期分析を行う。

この処理で求まったピッチ同期パワースペクトルを $|\hat{X}(k)|^2 \quad (0 \leqq k \leqq 256)$ とし、その平均パワー P を計算した後、逆スペクトル $|\hat{Y}(k)|^2$ を次の 3.12 式により求める。

$$|\hat{Y}(k)|^2 = P \cdot \frac{P}{|\hat{X}(k)|^2} = \frac{P^2}{|\hat{X}(k)|^2} \tag{3.12}$$

次に、$|\hat{Y}(k)|^2$ を次の 3.13 式の全極スペクトルでモデル化する。

$$|\hat{Y}(k)|^2 = \frac{G^2}{|1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2} \tag{3.13}$$

3.12 式と 3.13 式より、次の 3.14 式となる。

$$\begin{aligned} |\hat{X}(k)|^2 &= \frac{P^2}{G^2}|1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2 \\ &= EPG \cdot |1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2 \end{aligned} \tag{3.14}$$

この時、LPC 係数 β_l 及び利得 G は、まず、3.12 式で求まる逆スペクトル $|\hat{Y}(k)|^2$ から、次の 3.15 式で表される逆 DFT により定義される自己相関関数 r_l を逆 FFT 計算し、これを用いて LPC 分析を行うことにより求めることができる。

$$r_l = \frac{1}{N} \sum_{k=1}^{\frac{N}{2}} |\hat{Y}(k)|^2 \cos(\frac{2\pi}{N}kl) \qquad 0 \leqq l \leqq q \tag{3.15}$$

以上の分析側の処理により求まった LPC 係数 β_l 及び正規化パワー EPG をモデルパラメータとして出力して有声音源情報とする。

次に合成側においては 3.14 式の右辺第 2 項の逆スペクトルを β_l を 256 点 FFT することにより求めた後、3.14 式によりパワースペクトルを再生する。そしてこれより零位相で 256 点逆 FFT することにより求めたインパルス応答を、各ピーク位置に合わせて有声音残差を再生する。この時、3.1 式に示されるように 256 点のインパルス応答は、他のピッチ区間と重なってよい。以上の処理により各ピッチ周期毎に求まる、零点を有する励振パルスによる有声音源モデルを ZEP1(zero excitation pulse 1) と呼ぶことにする。

上記 ZEP1 において逆 LPC 係数 β_l の次数は、全周波数帯域の零点を近似できるだけの大きな値が必要である。これは振幅スペクトルにおいて特に中域以上の零点以外のピークが増えてしまい、最適な次数の設定が難しいためである。

そこで、$0 \leqq \omega \leqq \frac{\pi}{2}$ の音声パワーの大きな部分でのみ零点のモデル化を行い、$\frac{\pi}{2} \leqq \omega \leqq \pi$ の周波数域では正規化パワー EPG の平坦特性とするモデル 2 を考える。この場合、図 3.5 に示すように LPC 分析を行うための FFT のサイズは、ZEP1 の場合の $\frac{1}{2}$ でよい。以上の処理により求まる各ピッチ周期毎の有声音源モデルを、ZEP2 と呼ぶことにする。

3.4 有声音残差のモデル化実験

3.4.1 ZEP のモデル化特性

上記処理により求まる ZEP1 の振幅スペクトルの例を図 3.6 に示す。

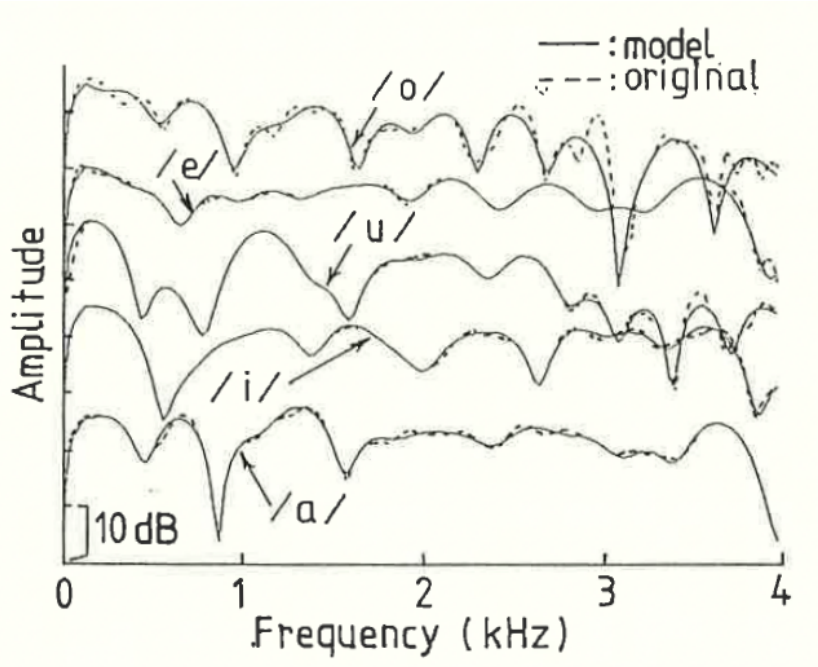


図 3.6 図 3.2(a) に対応する ZEP1 の振幅スペクトル

Fig. 3.6 Amplitude spectrum of ZEP1 correspondingto fig.3.2(a).

ここで、LPC 係数 β ι の次数 q は予備実験により 24 次に設定した。また直流分付近は近似しにくいため、直流分のパワーを f=125Hz の値の $\frac{1}{100}$ に定めその間はパワー領域で直線補間して再生した。図 3.6 より ZEP1 の振幅スペクトルは、有声音残差のピッチ同期振幅スペクトルの零点特性を非常に良く近似していることがわかる。

次に図 3.7 は、表 3.1 において用いた男女各 1 名の連続音声「あいうえお」について上記 ZEP1 を求め、その各ピッチ周期毎の振幅スペクトルからローカルな極小点を抽出し、そのヒストグラムを計算した結果である。

これより、ZEP1 による零点のモデル化が非常に安定して行えており、零点が正確に抽出されていることがわかる。また、零点周波数が音韻によらずかなり偏って分布しており、話者によってその分布が異なることがわかる。

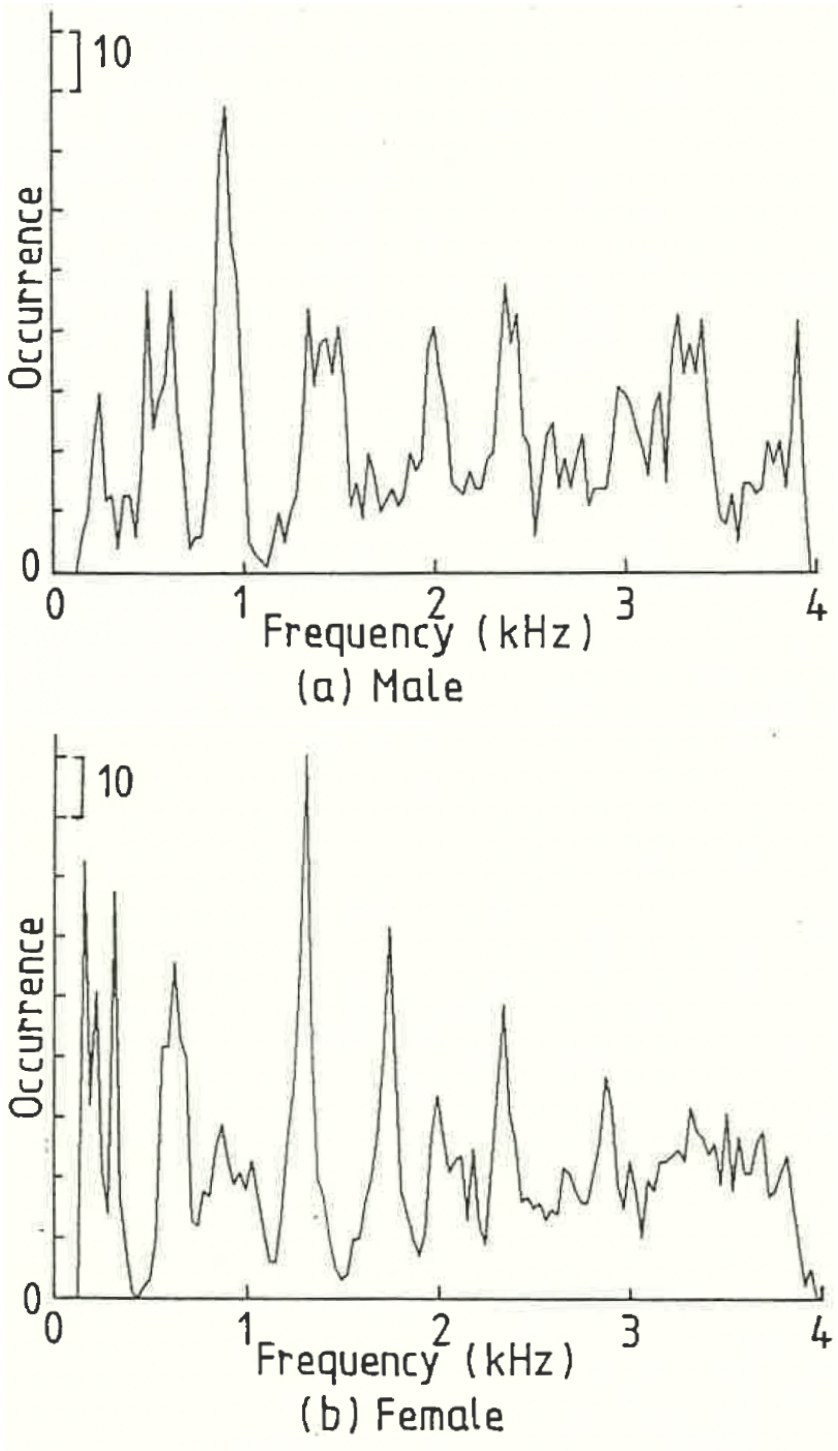


図 3.7 零点周波数の分布

Fig. 3.7 Distribution of zero frequency.

次に ZEP2 の振幅スペクトルの例を図 3.8 に示す。

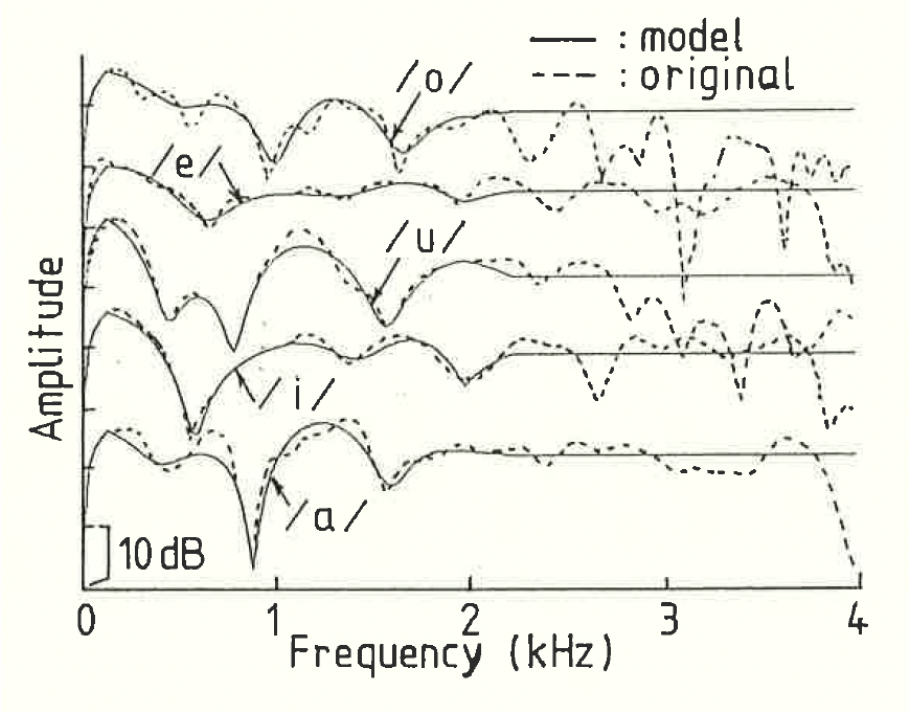


図 3.8 図 3.2(a) に対応する ZEP2 の振幅スペクトル

Fig. 3.8 Amplitude spectrum of ZEP2 correspondingto fig.3.2(a).

この場合の次数 q は予備実験により 8 次程度でよいことがわかった。また、直流分付近の処理は ZEP1 と同様であり、更に、$\omega = \frac{\pi}{2}$ 付近においてパワースペクトルが対数域で滑らかに接続されるようにパワー領域で直線補間を行った。図 3.8 より、$\omega = \frac{\pi}{2}$ の周波数帯域において、零点特性が十分に良く近似されていることがわかる。

図 3.9 に ZEP1 と ZEP2 の時間域波形の例を示す。

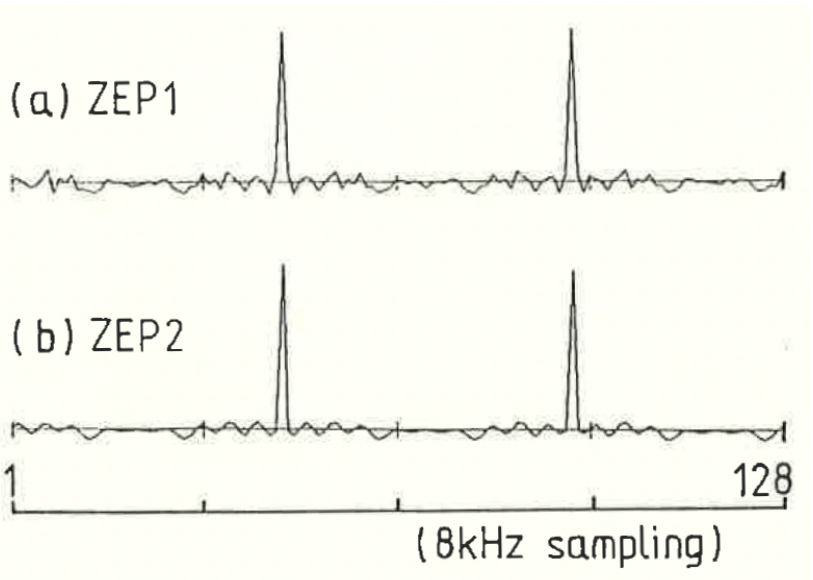


図 3.9 ZEP の時間域波形の例

Fig. 3.9 Examples of the time-domain waveform of ZEP.

3.4.2　ZEP を用いた LPC 合成音声の音質評価

有声音残差のピッチ同期分析によって求まる ZEP を有声音源とする合成音声の音質を評価するために、以下に示す 5 種類の音声の対比較試聴試験を行った。各対は順序を変えて 4 回ずつ出現するようにし、用いた原音声は男女各 1 名の 8KHz サンプリング、12bit 量子化による連続音声「あいうえお」である。そして音声 1 として、男女各原音声に対する 8 ビット logPCM 音声を用い、音声 2〜5 は、まず上記各原音声に対して、分析フレーム長 32msec、更新周期 5msec で 10 次 LSP 分析を行って得た有声音残差に対して、各々以下に示すモデル化を行い、再合成して得た合成音声である。なお、ピッチ開始点を視察により切り出した場合と、3.2.4 の表 3.2 の結果に基いてピーク位置から決定した場合については、合成音声の音質の差を知覚することはできなかった。

音声 2：各ピッチ周期毎にピーク位置を視察により抽出し、またピッチ開始点を各ピーク位置から表 3.2 の値を用いて求めた後、3.2.3 の方法で求めたピッチ同期振幅スペクトルからそのまま零位相でインパルス応答を求めて再合成した有声音残差。

音声 3：ZEP1 による有声音残差 (q=24)。

音声 4：ZEP2 による有声音残差 (q=8)。

音声 5：各ピッチ周期毎に視察により抽出したピーク位置にインパルスを立てて再合成した有声音残差。

ここで、今回はピッチ同期分析による ZEP1,2 の音質を直接評価するのが目的であるため、ZEP1,2 における各パラメータ β_l、EPG、又は β_l、EPG、及び音声合成時の LSP 係数の量子化は行っておらず、また各ピーク位置は視察により求めた。

被験者は男 9 名、女 1 名であり、各合成音声の音質の豊かさ、響き、滑らかさ、及び明瞭さなどに注意してもらいながらプレファレンススコアをとった。

図 3.10 に結果を示す。

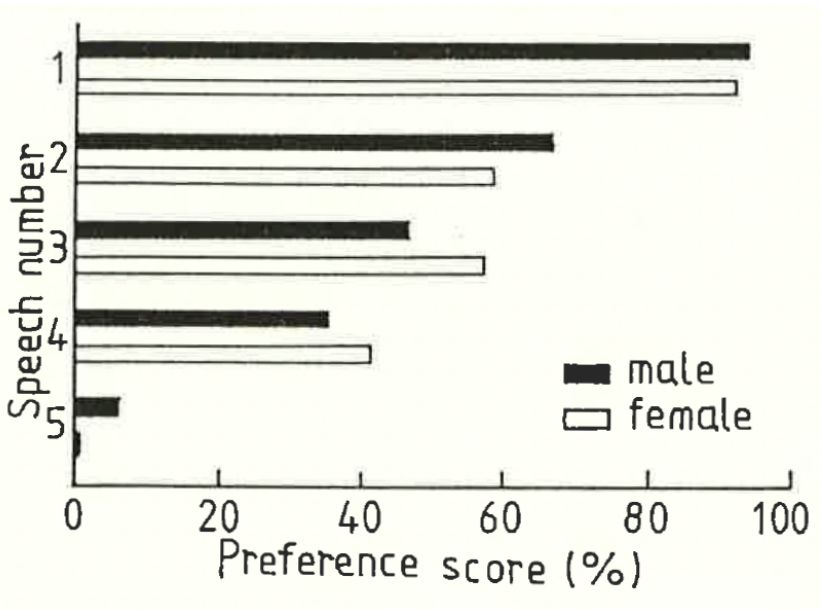


図 3.10 合成音声のプレファレンススコア

Fig. 3.10 Preference score of each synthesized speech.

音声 2,3,4 とも音声 1 の 8 ビット logPCM 音声には及ばなかったが、ピッチ同期振幅特性を付加した音源による合成音声の音質は、音声 5 の従来のインパルス列音源によるものよりかなり高いスコアを示した。また音声 2,3,4 の音質の差は特に女性において小さく、ZEP によるモデル化が有効に行えたことを示している。また音声 3,4 の音質の差は男女共に小さく、特に中帯域以下の零点特性が合成音声の音質に大きく寄与していることがわかる。この結果より、ZEP2 による有声音源パルスを用いれば、十分な音質の合成音声を得ることができることがわかった。

被験者の意見によると、音声 2 は音声 1 の音質にかなり近く滑らかであり、また音声 2,3,4 では、音声 5 に較べて音質が豊かに (又は太く) なり、響き、特に鼻にかかった感じも良く再生されており、更に、音声 5(特に女性) において顕著であった音韻「お」の不明瞭さがかなり改善されたという意見などが聞かれた。また、音声 2,3,4 の音質の差は、あまり感じられなかったということである。これらの結果は、零点特性が音声の個人性を表現する要素として重要であり、鼻にかかった母音などにおける音韻の明瞭度を向上させる効果も有するためであると考えられる。一方、滑らかさの点では特に音声 3,4 は音声 1 には僅かに及ばず、ぷつぷつという感じが残ったという意見が聞かれた。この傾向は特に女性において現れ、これはモデル化の誤差に加え、有声音残差のピッチ同期分析時に、各ピッチ周期の残差信号がその周期内で完全に減衰しきらず、得られたピッチ同期振幅スペクトルにも誤差が含まれたためであると思われる。

以上の結果より、有声音残差のピッチ同期振幅スペクトルは、30msec 程度の区間では準定常性を有することが確かめられたので、5〜30msec 程度のフレーム周期で有声音残差から代表的なピッチ周期を抽出し、ピッチ同期分析を実行してモデル化を行い、フレーム間で補間をすることにより大幅な情報量の圧縮が可能であると考えられる。この場合、残差のモデル化に必要な情報は、ZEP2 においては 8 次程度の LPC(又は LSP) 係数とパワー情報でよいため、従来のインパルス音源による LPC ボコーダに対して、 50%増し程度の情報量で実現することが可能である。また各ピッチ周期は、フレーム毎に代表的なピッチ区間が抽出できればよいため、第 2 章で提案したピーク抽出法等の他に、より簡単な方法でピーク抽出を行うことができると思われる。

以上のシステムの具体的な実現法及びその性能評価については、次の第 4 章で詳細に述べている。

更に図 3.7 の結果などから、有声音残差のピッチ同期分析から求まる零点特性は、話者の個人性などの情報も含んでいると考えられる。残差を用いた話者認識については、平均残差包絡に含まれる零点を利用した話者識別法 [44] が報告されているが、本手法における有声音残差のピッチ同期分析は、調波成分の影響を効果的に回避し零点特性を明確に観測できるという点で、話者認識への応用も有効であると考えられ、今後の検討課題である。

3.5 総括

第 3 章では、LPC 分析による有声音残差のピッチ同期分析を行うことにより、全極モデルとの差としての零点特性の正確な抽出を実現した。そして有声音残差のピッチ同期分析が、原音声の場合に較べて容易に行えること、及びそれにより得られたピッチ同期振幅スペクトルが、30msec 程度の区間では準定常性を有することを明らかにした。

また、その自動分析が可能であり、Ab-s などによる極零分析法に較べて、遥かに容易に零点を抽出できることを明らかにした。

上記分析法に基き、有声音源モデルとして零点を有する励振パルス ZEP を提案し、有声音残差の零点特性が正確にモデル化されることを明らかにした。

このモデルを用いて有声音残差を実際にモデル化し、それを用いた合成音声の音質の主観的な評価を行い、従来のインパルス音源に基く LPC 方式に比較して高品質な音声を合成できることを示し、特に、中帯域以下の零点特性のみをモデル化した ZEP2 によっても十分に良質な合成音声が得られ、低ビットレートのボコーダへの適用が可能であることを明らかにした。

第 4 章　LPC 有声音残差のピッチ同期メル逆 LSP 分析合成方式

4.1　概要

第 3 章において、LPC 有声音残差をピッチ同期分析することによって得たピッチ同期振幅スペクトルが極めて特徴的な零点特性を有することを示し、このピッチ同期振幅スペクトルに対して逆 LPC 分析を行うことによって得たパラメータを有声音源情報として伝送し、合成側ではこのパラメータを用いて零点を有する励振パルス (ZEP) を合成して有声音源とし LPC 合成を行う方式を提案した。

この方式により有声音残差のピッチ同期振幅スペクトルは、24 次程度の LPC 係数によって全周波数帯域が正確に近似できるが、これを用いて音質を損なわずに符号量を削減できる分析合成システムを実現する場合には次のような点に留意する必要がある。即ち、1. ピッチ同期振幅スペクトルは連続する数ピッチ区間では準定常的であるため、ピッチ区間を間引いて分析し補間を行うことにより符号量の圧縮が可能であるが、その場合補間特性の良いモデルパラメータをいかにして決定するか。2. モデルパラメータの補間及び励振パルスの合成を容易に行うための時間域でのフィルタの直接構成法をどう実現するか。3. 符号量削減のため音質の劣化を最小限に抑えつつ分析次数をいかにして低い次数に抑えるか等の問題を解決する必要がある。

本章では、逆 LPC 分析におけるモデルパラメータとして LSP 係数を用い、これを用いて構成した LSP 逆フィルタにより零点を有する励振パルス (ZEP) を時間域で直接合成できることを示して上記 1 及び 2 の問題に対処し、更に LPC 有声音残差のピッチ同期振幅スペクトルをメル化して逆 LSP 分析を行うことによって得たメル LSP 係数を用いることにより分析次数が 12 次程度で良好なモデル化が行えることを示し、メル LSP 係数を用いた LSP 逆フィルタの直接構成法を提案して上記 3 の問題に対処している。

以上の分析合成系により合成された有声音残差について、メル領域でのピッチ同期振幅スペクトルの平均スペクトル歪を評価尺度として、メル逆 LSP のモデル化次数、量子化特性、及び補間特性について検討し、各々の最適化を図った。

この結果に基いて構成される有声音残差のピッチ同期メル逆 LSP 分析合成システムによる合成音声の主観的音質評価を行い、4.8Kbits/sec 程度のビットレートで高品質な音声を合成でき、低ビットレートボコーダへの応用が可能であることを示す [45]。

4.2　有声音残差のピッチ同期メル逆 LSP 分析合成系

4.2.1　有声音残差のピッチ同期逆 LPC に基く有声音残差の合成

第 3 章の 3.3 節において、特徴的な零点特性を有する有声音残差のピッチ同期スペクトル $|\hat{X}(k)|^2$ は、3.12 式によりその逆スペクトル $|\hat{Y}(k)|^2$ を求めた後、3.15 式によって逆スペクトルに対応する自己相関関数 r_l を求め、LPC 分析を行うことによって得た LPC 係数 β_l 及び正規化パワー EPG によりモデル化できることを示した。従って、この β_l 及び EPG から有声音残差を合成

するためには、原理的には 3.3 節で述べたように、上記と逆の処理によりピッチ同期スペクトル $|\hat{X}(k)|^2$ を計算し、逆 FFT によりそのインパルス応答を計算して合成することができる。しかし実際には、ピッチ同期スペクトル $|\hat{X}(k)|^2$ は、LPC 係数 β_l 及び正規化パワー EPG を用いた全極モデルによって前記 3.14 式のように表現できる。そして同式の右辺第 2 項は、残差を求めるための LPC 逆フィルタの周波数特性そのものであるため、LPC 係数 β_l 及び正規化パワー EPG から 3.14 式のインパルス応答を 1.6 式の LPC 逆フィルタを用いて、次の 4.1 式として求め、これを各ピーク位置に合わせて有声音残差を再合成できる。

$$\hat{h}_e(m) = \delta(m) - \sum_{i=1}^{q} \beta_i \delta(m-i) \tag{4.1}$$

$$\text{但し、} \begin{cases} \delta(m) = EPG & for\, m = 0 \\ \delta(m) = 0 & for\, m \neq 0 \end{cases}$$

ここでインパルス応答 $h_e(m)$ は最小位相特性を有し原残差の位相特性に近い [15]。

4.2.2 逆 LPC 分析合成系の LSP 化

逆 LPC 分析によって求まる LPC 係数を符号化して伝送する場合、係数の量子化及び補間を行う必要がある。しかし LPC 係数は符号化特性がそれほど良くないため、補間特性及び量子化特性の良い LSP 係数に変換して伝送、合成することが考えられる [26]。4.2.1 の有声音ピッチ同期逆 LPC 分析に LSP 係数を導入する場合、分析側では逆 LPC 分析によって求まった LPC 係数から求めることができ、合成側では LSP 逆フィルタを構成して振幅が EPG のインパルスを入力することにより各ピッチ周期毎の有声音残差を直接合成することができる。ここで LSP 逆フィルタの伝達関数 $A_q(z)$ は、次の 4.2 式によって求めることができる [26]。

$$\begin{aligned} A_q(z) = \frac{z^{-1}}{2}[& \sum_{i=2(i=even)}^{q} (c_i + z^{-1}) \prod_{j=0(j=even)}^{i-2} (1 + c_j z^{-1} + z^{-2}) \\ & - \prod_{j=2(j=even)}^{q} (1 + c_j z^{-1} + z^{-2}) \\ & + \sum_{i=1(i=odd)}^{q-1} (c_i + z^{-1}) \prod_{j=-1(j=odd)}^{i-2} (1 + c_j z^{-1} + z^{-2}) \\ & - \prod_{j=1(j=odd)}^{q-1} (1 + c_j z^{-1} + z^{-2})] + 1 \end{aligned} \tag{4.2}$$

$$\text{但し、} \begin{cases} c_i = -2\cos(2\pi f_i) & 1 \leqq i \leqq q \\ c_0 = c_{-1} = -z^{-1} & \\ f_i(\text{Hz})\text{はLSP係数} & \end{cases}$$

4.2.3 逆 LPC 分析のメル LSP 表現

ピッチ同期有声音残差を時間域で合成するためには、第 3 章 3.3 節の 3.14 式におけるピッチ同期振幅スペクトルを全周波数帯域でモデル化する必要があるが、分析時において全周波数帯域を正しくモデル化するために必要な次数は、第 3 章 3.4.1 で示したように 24 次程度になってしまう。そこで人間の聴覚の周波数特性がメル目盛りに従うことを利用して、逆 LPC 分析時における LPC 係数をメル LSP 係数により表現し次数の削減を行う。

人間の聴覚の周波数特性は高域になるほど鈍感になることが知られており [46]、特に聴覚心理に基くメル目盛によってその特性が良く表現される。メル目盛は、次の 4.3、4.4、及び 4.5 式で表される全域通過フィルタによる直線目盛から非直線目盛への周波数変換において、8KHz で a=0.31 と選ぶことによって得ることができる [47][48]。

$$\tilde{z}^{-1} = \frac{(z^{-1} - a)}{(1 - az^{-1})} \tag{4.3}$$

$$\tilde{\omega} = \omega + 2\tan^{-1}\frac{a\sin\omega}{1 - a\cos\omega} \tag{4.4}$$

$$\frac{\mathrm{d}\tilde{\omega}}{\mathrm{d}\omega} = \frac{1 - a^2}{1 + a^2 - 2a\cos\omega} \tag{4.5}$$

これを利用して、第 3 章 3.3 節に基く有声音残差のピッチ同期逆 LPC 分析のメル化を行う。ここでは 3.3 節 3.15 式の自己相関関数 r_l を求めるための逆 DFT の計算を、次の 4.6、4.7 式に示すメル領域の逆 DFT[47] で置き換えて得たメル自己相関関数 $\tilde{r}_l$ に対して、通常の LPC 分析、LSP 分析を行うことによりメル LSP 係数を求める。

$$\tilde{r}_i = \frac{1}{N}\sum_{k=0}^{\frac{N}{2}} |\hat{Y}(k)|^2 U_{lk} \qquad \text{但し、} 0 \leqq l \leqq q \tag{4.6}$$

$$U_{lk} = \frac{\mathrm{d}\tilde{\omega}_k}{\mathrm{d}\omega_k}\cos(l\tilde{\omega}_k) \tag{4.7}$$

ここでメル逆 DFT の計算は、4.7 式に $\omega_k = \frac{2\pi k}{N}$ 及び 4.4 式、4.5 式を代入すると、次の 4.8 式となる。

$$U_{lk} = \frac{1 - a^2}{1 + a^2 - 2a\cos\frac{2\pi}{N}k} cos(l(\frac{2\pi}{N}k + 2\tan^{-1}\frac{a\sin\frac{2\pi}{N}k}{1 - acos\frac{2\pi}{N}k})) \tag{4.8}$$

従って $k = 0$〜$\frac{N}{2}$、$l = 0$〜q で 4.8 式を計算し、予めテーブルを作成しておくことにより 4.6 式を高速に計算することが可能である。

4.2.4 LSP 逆フィルタのメル化

逆 LPC 分析によって求まる LSP 係数から各ピッチ周期毎の有声音残差を再合成するためには、4.2.2 で述べたように 4.2 式の伝達関数を有する LSP 逆フィルタのインパルス応答を求めれば良いため、4.2.3 で述べた処理で求まるメル LSP 係数を直接入力とするメル LSP 逆フィルタを構成することを考える。

LSP 逆フィルタの伝達関数には 4.2 式に示すように帰還ループが含まれないため、同式の各コンポーネント $\frac{z^{-1}}{2}$、$c_i + z^{-1}$、$1 + c_j z^{-1} + z^{-2}$ における z^{-1} を 4.3 式で表される $\tilde{z}^{-1}$ に置き換えれば良い。今、$H0(z) = \frac{\tilde{z}^{-1}}{2}$、$HA_i(z) = \tilde{c}_i + \tilde{z}^{-1}$、$HB_j(z) = 1 + \tilde{c}_j\tilde{z}^{-1} + \tilde{z}^{-2}$、$\tilde{c}_i = -2\cos(2\pi\tilde{f}_i)(\tilde{f}_i(Hz)$はメル$LSP$係数)、$\tilde{c}_0 = \tilde{c}_{-1} = -\tilde{z}^{-1}$ とすれば 4.3 式より、次の 4.9、4.10、及び 4.11 式と計算できる。

$$H0(z) = \frac{\tilde{z}^{-1}}{2} = \frac{1}{2}\frac{-a + z^{-1}}{1 - az^{-1}} \tag{4.9}$$

$$HA_i(z) = \tilde{c}_i + \tilde{z}^{-1} = \frac{(\tilde{c}_i - a) + (1 - \tilde{c}_i a)z^{-1}}{1 - az^{-1}} \tag{4.10}$$

$$HB_j(z) = 1 + \tilde{c}_j + \tilde{z}^{-1} + \tilde{z}^{-2} = \frac{(1 - \tilde{c}_j a + a^2) + (\tilde{c}_j - 4a + \tilde{c}_j a^2)z^{-1} + (1 - \tilde{c}_j a + a^2)z^{-2}}{1 - 2az^{-1} + az^{-2}} \tag{4.11}$$

上記 3 つの式と 4.2 式よりメル LSP 逆フィルタの伝達関数 $A_q(z)$ は、次の 4.12 式によって決定することができる。

$$\begin{aligned} A_q(z) = H0(z)[& \sum_{i=2(i=even)}^{q} HA_i(z) \prod_{j=0(j=even)}^{i-2} HB_j(z) - \prod_{j=2(j=even)}^{q} HB_j(z) \\ & + \sum_{i=1(i=odd)}^{q-1} HA_i(z) \prod_{j=-1(j=odd)}^{i-2} HB_j(z) - \prod_{j=-1(j=odd)}^{q-1} HB_j(z)] + 1 \end{aligned} \tag{4.12}$$

但し、$HB_0 = HB_{-1} = 0$

図 4.1 にサンプリング周波数 8KHz、分析次数が偶数次の時のメル LSP 逆フィルタの構成を示す。

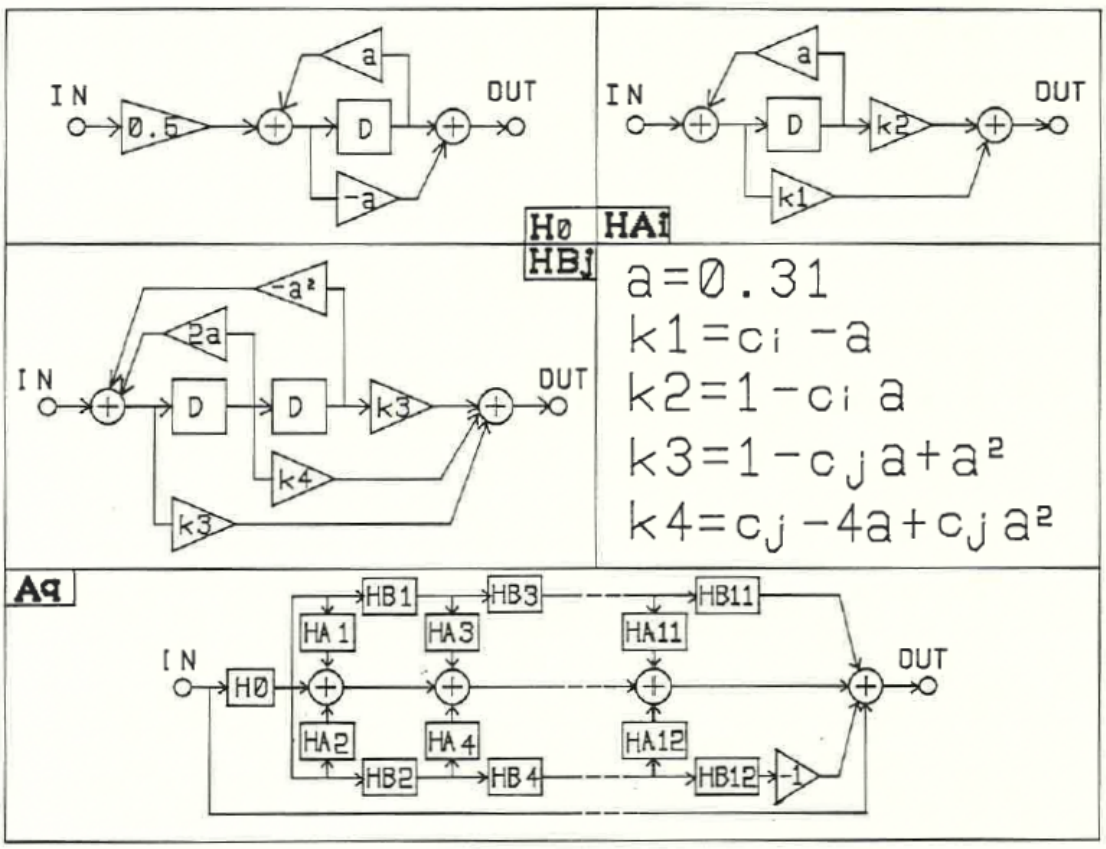


図 4.1 メル LSP 逆フィルタ

Fig. 4.1 Mel LSP inverse filter.

4.2.5 ピッチ同期メル逆 LSP 分析合成手順

以上の各原理に基く有声音残差のピッチ同期メル逆 LSP 分析と合成の手順について以下に簡単に記す。

(分析側の処理手順)

(1) 始めに有声音残差 ϵ_n から各ピッチ区間を切り出し 0 を付加して N 点とし、N 点 FFT によりピッチ同期パワースペクトル $|\hat{X}(k)|^2 \quad (0 \leqq k \leqq N)$ とその平均パワー P を計算する。ここでピッチ間隔は 8KHz 標本化音声では 128 サンプルを越えることはまれなため、N=128 として

問題ない。

(2) 次に第 3 章 3.3 節の (3-12) 式で逆スペクトル $|\hat{Y}(k)|^2$ を求めた後、メル逆 DFT によってメル自己相関関数 $\tilde{r}_l \quad (0 \leqq l \leqq q)$ を求め、LPC 分析、LSP 分析によりメル LSP 係数 $\tilde{f}_l \quad (1 \leqq l \leqq q)$ を計算する。

(3) 一方、平均パワー P と LPC 分析時に求まる利得 G(3.3 節 3.13 式) から、同節 3.14 式によって正規化パワー EPG の平方根 $\sqrt{EPG}$ を振幅情報として計算する。

(4) 以上、分析側で求まる各フレーム毎のメル LSP 係数 $\tilde{f}_l$、振幅情報 $\sqrt{EPG}$、及び各ピッチ周期のピーク位置を出力する。

(合成側の処理手順)

(5) 分析側から伝送されてきたメル LSP 係数 $\tilde{f}_l$ を用いて 4.9〜4.12 式 (図 4.1) のメル LSP 逆フィルタを構成し、振幅 $\sqrt{EPG}$ のインパルスを入力することによって各ピッチ周期毎のインパルス応答を計算し、各ピーク位置に合わせて有声音残差 $\hat{\epsilon}_n$ を合成する。

4.3 ピッチ同期メル逆 LSP の諸特性

本節では、メル逆 LSP のモデル化次数、量子化特性、及び補間特性について検討する。

4.3.1 メル逆 LSP のモデル化次数

4.2 節の分析合成系における各分析次数毎のモデル化特性を図 4.2 に示す。

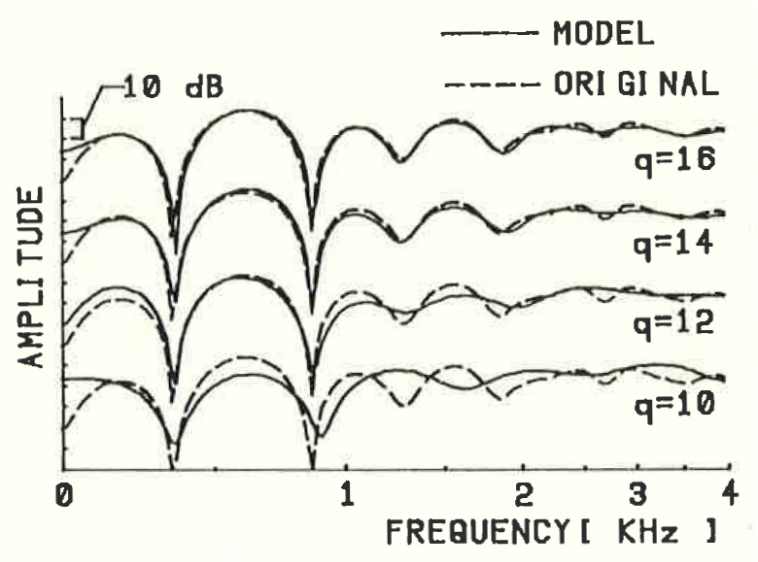


図 4.2 各次数毎のモデル化特性

Fig. 4.2 Characteristic of modeling at each order.

この図において、破線は 1 ピッチ周期の有声音残差のピッチ同期振幅スペクトルをメル周波数目盛に変換して示したものであり、実線は前節の分析合成系によって合成された各分析次数毎のインパルス応答の振幅スペクトルをメル目盛上で示したものである。同図より次数 q が 10 次では零点を良く近似しきれてないが、12 次では低域の Q の大きな零点はほぼ良好に近似している。また 14 次、16 次では全域にわたって良好な近似が行われている。次に次数が 14 次の時の合成された有声音残差を図 4.3 に示す。

図 4.3 原有声音残差と合成有声音残差

Fig. 4.3 Original and synthesized voiced residuals(q=14).

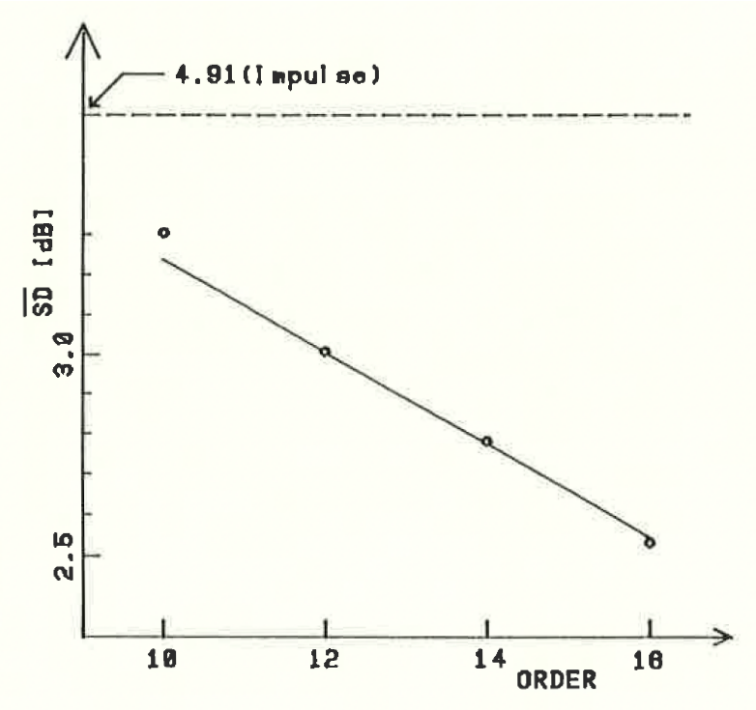


図 4.4 モデル化次数と平均スペクトル歪の関係

Fig. 4.4 Relationship between mean spectral distortion and modeling order.

次に図 4.4 は、男女各 1 名の約 2 秒の短文章「横浜は何県にありますか」を 8KHz サンプリング、フレーム周期 5msec、分析次数 10 次で LSP 分析して求まる有声音残差の各ピッチ周期を切り出してメル化した後、128 点 FFT により計算したメル対数振幅スペクトル $PZ_0(k)$ と、前節の分析合成系によって合成された各ピッチ周期毎のインパルス応答をメル化した後、128 点 FFT によ

り計算したメル対数振幅スペクトル $PZ_1(k)$ とを用いてスペクトル歪 SD を次の 4.13 式により計算し、これを全ピッチ周期において平均して求めた平均スペクトル歪 SD の結果を分析次数毎に示したものである。

$$SD = \sqrt{\frac{1}{128}\sum_{k=1}^{64}(PZ_o(k) - PZ_1(k))^2} \tag{4.13}$$

ここでは各パラメータの量子化と補間は行っておらず、また各ピッチ区間は各ピーク位置を視察により抽出した後、隣接ピーク間の 42.3%の位置をピッチ開始点として切り出している [45]。更に $PZ_0(k)$ と $PZ_1(k)$ は互いに平均振幅が等しくなるように正規化しておく。なお図 4.3 において、各ピッチ周期毎にモデル化を行わずフラットスペクトルとした場合の SD を図 4.4 の破線で示す。同図より、SD は次数が増えるに従って直線的に減少するが、10 次ではその直線性が若干失われてる。

図 4.2 と図 4.4 より、人間の低域重視という聴覚特性 [46] を考慮すると有声音残差のピッチ同期振幅スペクトルは、12〜14 次程度で良好に近似でき、前章で述べた逆 LPC 分析で次数が 24 次程度必要であったのに比較して、10 次程度の分析次数の削減になる。

4.3.2 メル逆 LSP の量子化特性

図 4.5 に、分析次数が 12 次の場合について、図 4.4 と同一条件で有声音残差のピッチ同期メル逆 LSP 分析を行って得られるメル LSP 係数の存在範囲と度数分布を示す。

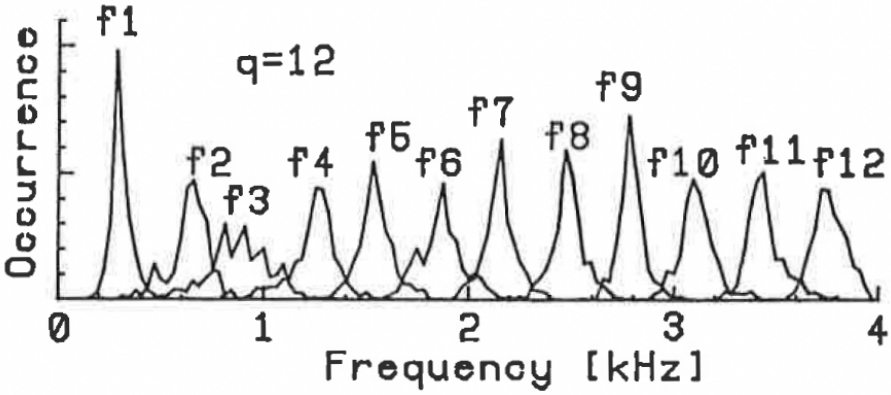


図 4.5 メル LSP 係数の度数分布

Fig. 4.5 Distribution of mel LSP coefficiences(q=12).

この図より各次数毎のパラメータの存在範囲はほぼ等間隔であり、その存在幅は最大 800Hz 程度である。従って各次数毎にその存在範囲内で量子化を行うことによって、効率の良い量子化を行える [26]。

次に図 4.6 は、図 4.4 と同一条件でメル逆 LSP 分析を行って得られるメル LSP 係数を用いて、まず量子化をせずに合成を行って得た各ピッチ周期毎のインパルス応答から計算したメル対数振幅スペクトル $PZ_1(k)$ と、各次数毎のメル LSP 係数を図 4.5 の存在範囲内で同一の量子化ビット数で量子化し合成を行なって計算したメル対数振幅スペクトル $PZ_2(K)$ とを用いて、図 4.4 と同様に計算した平均スペクトル歪 $\bar{SD}$ の結果を、分析次数が 12 次の場合について量子化ビット数毎に示したものである。

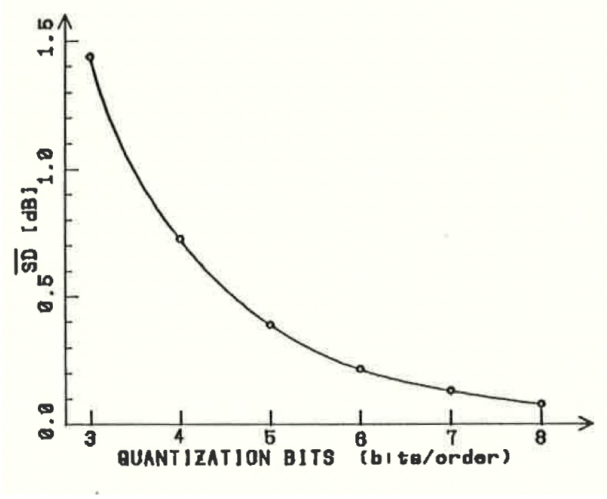


図 4.6 各次数毎の量子化ビット数と平均スペクトル歪の関係

Fig. 4.6 Relationship between mean spectral distortion and bit rate per order(q=12).

この図より各次数あたり 4 ビット程度で量子化すれば、$\bar{SD}$ は十分に小さくできる。

4.3.3 メル逆 LSP の補間特性

第 3 章で述べたように、有声音残差のピッチ同期メル逆 LSP 分析合成を行う場合、連続するピッチ周期ではピッチ同期振幅スペクトルは準定常的であるため、一定のフレーム間隔毎に代表的なピッチ区間を抽出しピッチ区間を間引いて分析・符号化を行うと共に、そのフレームの平均ピッチ周期を符号化し、合成側では平均ピッチ周期からインパルス列を生成した後、メル LSP 係数をフレーム間で補間し各インパルス位置毎の係数を計算して合成すれば、符号量の圧縮が可能であると考えられる。このことを、メル逆 LSP の補間特性を調べることにより確認する。

ここで、有声音残差の各ピッチ周期毎のメル対数振幅スペクトルをリファレンスとし、これに対してピッチ周期を一定間隔毎に間引いて分析し、その間の各ピッチ周期毎のメル対数振幅スペクトルを補間して求め、リファレンスに対する平均スペクトル歪を計算したものを、メル逆 LSP の補間特性として定義する。

リファレンスのスペクトルは、通常、分析間隔を十分に短くしたものを用いるが、有声音残差のピッチ同期分析では分析間隔をピッチ周期より短くすることはできないため、図 4.4 の場合と同一の条件で、有声音残差の各ピッチ周期毎にメル逆 LSP 分析・合成を行って得たメル対数振幅スペクトル $PZ_1(k)$ をリファレンスとして用いる。

一方、補間をするために一定間隔でピッチ周期を間引く場合にも、ピッチ周期単位で行うことになるため、補間によるスペクトルは次のようにして求める。まず補間間隔に対応するフレーム毎に、その中央位置に最も近いピッチ周期を切り出してメル逆 LSP 分析を行い、得られたメル LSP 係数を各フレームの代表値とする。これよりその係数をフレーム間で直線補間し、上記リファレンスの各ピッチ周期のピーク位置に対応するメル LSP 係数を計算して合成を行い、メル対数振幅スペクトル $PZ_3(k)$ として求める。

以上の処理で求まった $PZ_1(k)$ 及び $PZ_3(k)$ から、平均スペクトル歪 $\bar{SD}$ を図 4.4 の場合と同様に計算し、上記補間間隔での補間特性とする。

図 4.7 に、分析次数が 12 次、補間 (フレーム) 間隔が 10、20、及び 30msec の各場合について $\bar{SD}$ を求めた結果を示す。

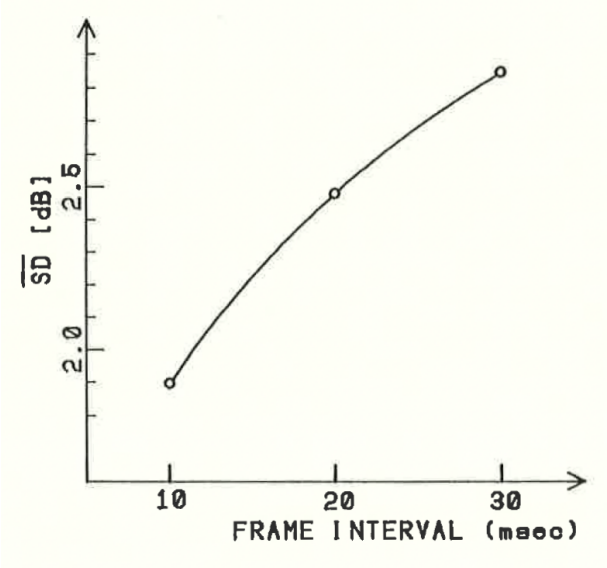


図 4.7 フレーム補間間隔と平均スペクトル歪の関係

Fig. 4.7 Relationship between mean spectral distortion and frame interval of interpolation(q=11).

この図の結果は、メル LSP 係数の補間特性というよりは、分析するピッチ区間を間引いた時に分析されなかったピッチ区間の特性を補間によってどの程度再現できるかという、一種の有声音残差のピッチ同期振幅スペクトルの定常性を表しているといえる。先の手法において、ピッチの切り出し位置の誤差によりピッチ同期振幅スペクトルの平均スペクトル歪が 2.4dB 程度の時に、合成音声の音質の差を知覚できなかったことを考慮して図 4.7 を検討すると、補間間隔を 10 又は 20msec 程度に設定すれば効率の良い符号化が行えると考えられる。

4.4 ピッチ同期メル逆 LSP 分析合成システム

以上、4.2 節の分析合成原理及び 4.3 節の諸特性の実験結果に基く有声音残差のピッチ同期メル逆 LSP 分析合成システムの構成を図 4.8 に示す。

始めに一定のフレーム間隔 (10〜20msec) 毎に平均ピッチ間隔 $T_p(m)$ を抽出した後、各フレーム m 毎の有声音残差 $\epsilon_n(m)$ からそのフレームの中央位置に最も近いピッチ区間を代表ピッチ区間として切り出し、4.2.5 で述べた手順によりピッチ同期パワースペクトル $|\hat{X}_k(m)|^2$、逆スペクトル $|\hat{Y}_k(m)|^2$、メル自己相関関数 $\tilde{r}_l(m)$ を介してメル LSP 係数 $\tilde{f}_l(m)$ $(0 \leqq l \leqq q)$ を計算する。次に振幅情報は、フレーム間の補間特性を良くするために $\sqrt{EPG}$ ではなく、各フレーム毎の $\epsilon_n(m)$ の平均パワーの平方根 $AMP(m)$ とする。以上、分析側で求まる各フレーム m 毎のメル LSP 係数 $\tilde{f}_l(m)$、振幅情報 $AMP(m)$、及び平均ピッチ間隔 $T_p(m)$ を符号化する。

合成側においては分析側から伝送されてきた各情報を復号化した後、まず、平均ピッチ間隔 $T'_p(m)$ 及び振幅情報 $AMP'(m)$ をフレーム間で補間しながら各フレーム毎のインパルス列 $\zeta_n(m)$ を作成する。次にメル LSP 係数 $\tilde{f}_l(m)$ をフレーム間で補間してインパルス列 $\zeta_n(m)$ の各インパルス位置毎の係数を求め、メル LSP 逆フィルタを構成し (図 4.1)、インパルス列 $\zeta_n(m)$ を入力することによって有声音残差 $\hat{\epsilon}_n(m)$ を合成する。

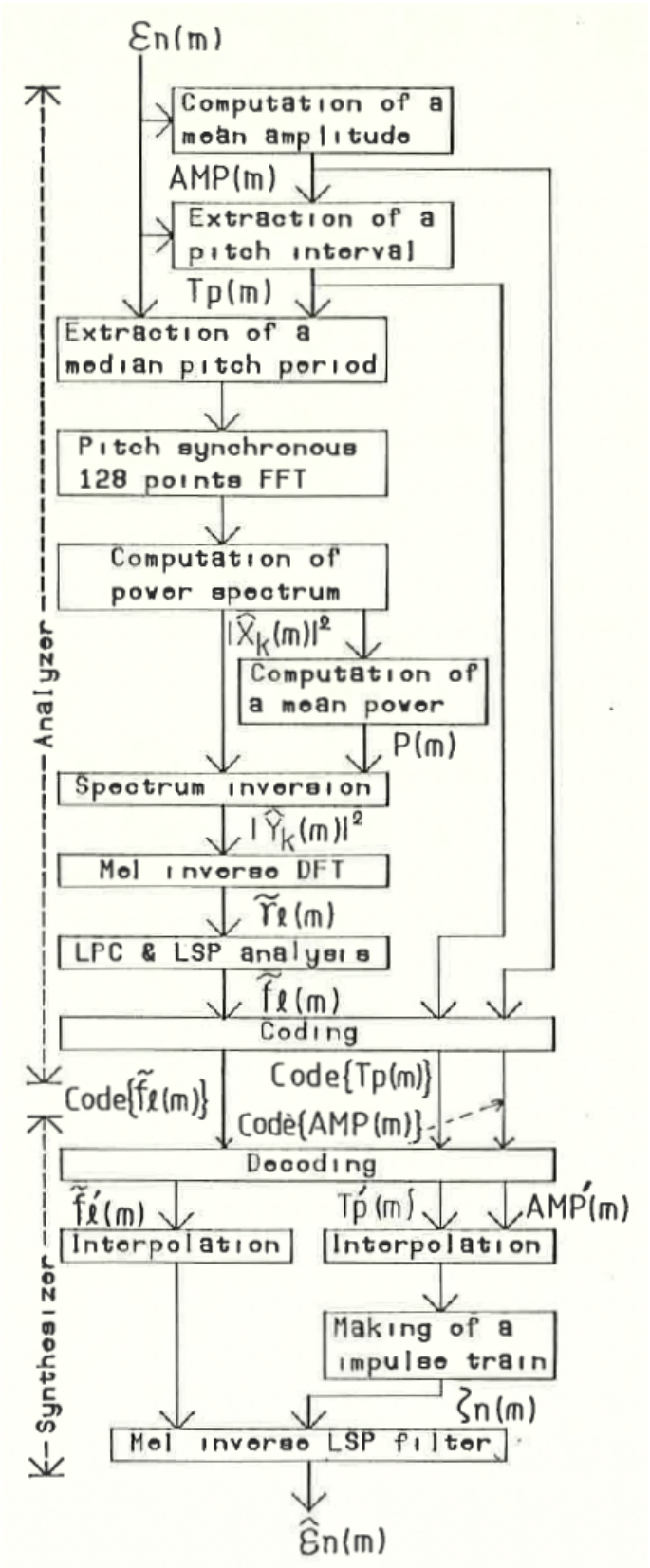


図 4.8 LPC 有声音残差のピッチ同期メル逆 LSP 分析合成システム

Fig. 4.8 Block diagram of the pitchsynchronous mel inverse LSP analysis-synthesis system.

4.5 合成音声の音質評価

図 4.8 のシステムにより合成される有声音残差を有声音源とする合成音声の音質を評価するために、表 4.1 の各音声に対して平均オピニオン値 (MOS)[49] によるオピニオン試聴試験を行った。

表 4.1 音質評価に用いた合成音声

Table 4.1 Synthesized voices used for sound quality estimation.

<table>
<tr><td colspan="2">音声</td><td>1 (本方式)</td><td>2</td><td>3</td></tr>
<tr><td colspan="2">声道特性</td><td colspan="2">10 次 LSP　37bits/frame</td><td rowspan="6">log PCM</td></tr>
<tr><td rowspan="3">有声音源</td><td>ピッチ</td><td colspan="2">7bits/frame</td></tr>
<tr><td>振幅</td><td colspan="2">7bits/frame</td></tr>
<tr><td>係数</td><td>12 次メル LSP 係数　45bits/frame</td><td>-</td></tr>
<tr><td colspan="2">無声音源</td><td colspan="2">ホワイトノイズ</td></tr>
<tr><td colspan="2">更新周期 (msec)</td><td>20</td><td>5</td></tr>
<tr><td colspan="2">情報量 (Kbits/s)</td><td>4.8</td><td>10.2</td></tr>
</table>

ここで用いた原音声は、男女各 1 名の 8KHz サンプリング、12bit 量子化による約 2 秒の短文章「横浜は何県にありますか」であり、表 4.1 の音声 1 は図 4.8 のシステムにより合成される有声音残差を有声音源とする本方式による合成音声、リファレンスとして用いた音声は、インパルス列を有声音源とする LSP 合成音声 (表 4.1 の音声 2) 、及び 4〜8 ビット logPCM 音声 (表 4.1 の音声 3) である。また被験者は男女 10 名である。

表 4.2 に、各音声に対する MOS 値とその標準偏差 σ を示す。

表 4.2 MOS 値による音質評価結果

Table 4.2 Result of sound quality estimation by MOS values.

音声サンプル	MOS	σ
本方式 (4.8Kbits/s)	1.4	0.6
従来方式 (10.2Kbits/s)	0.5	0.4
8 ビット logPCM	3.0	0.4
7 ビット logPCM	2.5	0.6
6 ビット logPCM	1.9	0.5
5 ビット logPCM	1.3	0.5
4 ビット logPCM	0.5	0.3

この試験の結果、4.8Kbits/sec における本方式による合成音声は、インパルス列を有声音源とする 9.6Kbits/sec 強における LSP 合成音声よりも高品質であり、5 ビット logPCM 音声の音質に相当することがわかる。5 ビット logPCM 音声は、40Kbits/sec のビットレートであるため約 1/10 の圧縮率であり、大きな情報圧縮が可能であるという結果が得られた。また、インパルス列を有声音源とする LSP 合成音声は、4.8Kbits/sec 以上でその音質が飽和することが報告されている [26] ことを考えると、本方式は従来の LSP 合成音声の音質の飽和限界を高める方式であるといえる。試聴者によればインパルス列を有声音源とする LSP 合成音声に比較してノイズ感が減少して音質が豊かになり、内容を聞き取りやすかったという意見が聞かれた。ただし、無声音部及び無

声音部と有声音部の接続部は聞き取りにくかったという意見も聞かれた。これは本方式がボコーダとしての構成を有し有声無声判別を行っており、無声音部は従来と同様にホワイトノイズで励振したことにより音質劣化が生じたためであると考えられ今後の課題である。

4.6 総括

第 4 章では、LPC 分析による有声音残差の新しい符号化法として、ピッチ同期メル逆 LSP 分析合成方式を提案した。その結果、零点特性を有する有声音残差のピッチ同期振幅スペクトルを、メル領域の逆 LSP 分析によって求まるメル LSP 係数を用いて、12 次程度の分析次数で高能率に符号化できることを示した。更に、メル LSP 係数から有声音残差を時間域で直接合成するメル LSP 逆フィルタを提案し、有声音残差を容易に合成できることを明らかにした。

次に、合成された有声音残差についてメル領域でのピッチ同期振幅スペクトルの平均スペクトル歪を評価することにより、メル逆 LSP のモデル化次数、量子化特性、及び補間特性について検討し、各々の最適化を図った。

最後に合成音声の主観的な音質評価を行い、4.8Kbits/sec 程度のビットレートで高音質な音声を合成できることを示し、低ビットレートのボコーダへの応用が可能であることを明らかにした。

第 5 章　結　　論

本研究は、低・中ビットレートでの音声生成モデルに基く LPC 分析合成方式の音声分析合成システムにおける音源情報の正確な抽出、並びに音源生成モデルに基く音源の新しいモデル化と、それを用いた分析合成系の高音質・高能率化を達成することを目的とし、マクロな音源情報抽出法として LPC 残差のピッチピークの位置と振幅を正確に求める方式について研究を行い、更に音源の生成モデルに基く LPC 有声音残差からのミクロな音源情報抽出とそのモデル化について研究を行ったものである。

緒論においては、音声の高能率符号化の意義と本研究の位置づけとの関係、及び本研究の目的と概要について述べ、音声分析合成システムにおける音源情報の抽出とモデル化に関する研究であることを明らかにした。

第 1 章では、音声生成モデル、特に LPC 方式に基く音声合成と残差からの音源情報抽出とモデル化について概観し、音源情報抽出とモデル化に関する従来の諸研究と問題点について述べ、本研究の意義を明らかにした。

第 2 章では、LPC 方式における音源情報抽出に関する最も基本的な問題である残差からのマクロな音源情報抽出、即ち各ピッチピークの正確な抽出法について述べた。特に、ピッチピーク以外の成分を含む残差からピッチ周期のみを正確に抽出する手法として、時間域信号を周波数スペクトルとみなし、その振幅包絡特性を LPC 分析を応用した全極モデルとして求めることにより、また鼻音部分を識別し、その部分の処理を適切に行うことにより、ピッチピークの位置と振幅を正確に求めることが可能であることを示した。そしてこのシステムから求まる音源情報を音声分析合成系に適用した結果、無声音と有声音のわたり部分の情報やピッチ周期の微妙な変化を正確にモデル化でき、合成音声の音質が向上することを明らかにした。

第 3 章では、LPC 方式における有声音残差から、マクロな音源情報の抽出に加えてミクロな音源情報を抽出する手法について提案した。第 2 章では、LPC 音声分析合成システムのためのマクロな音源情報の正確な抽出を実現し、1.2kbits/sec〜4.8kbits/sec の低ビットレートの伝送帯域では十分に高能率な音声符号化の達成が可能であることを示唆した。その一方では、第 1 章において 4.8Kbits/sec〜9.6kbits/sec の低・中ビットレートの音声符号化に LPC 方式を利用しようとした場合、有声音残差信号からミクロな音源情報として全極モデルとの差としての零点特性を抽出し、モデル化することが必要であることを述べた。本手法はそのような要請に答えたものであり、そのための手法として有声音残差のピッチ同期分析が有効であり、それにより求まるピッチ同期振幅スペクトルが極めて特徴的な零点特性を有することを示し、同分析が原音声の場合に較べて容易に行えること、ピッチ同期振幅スペクトルが 30msec 程度の区間では準定常性を示し、その自動分析が可能であることを明らかにした。この分析法に基き、零点を有する励振パルス ZEP を提案し、その有効性を確認した。また、このモデルを用いて有声音残差を実際にモデル化し、それを用いた合成音声の音質の主観的な評価を行って、従来のインパルス音源による LPC 方式に比較して高品質な音声を合成できることを明らかにした。

第 4 章では、第 3 章で提案した LPC 有声音残差のピッチ同期分析を基にして、それを 4.8kbits/sec 程度の低ビットレートの音声符号化方式に適用する方式として、有声音残差のピッチ同期振幅スペクトルの新しい符号化法—ピッチ同期メル逆 LSP 分析合成方式を提案し、特徴的な

零点特性を有する上記スペクトルをメル領域の逆 LSP 分析によって求まるメル LSP 係数を用いて、12 次程度の分析次数で正しくモデル化できることを示した。更に、メル LSP 係数から有声音残差を時間域で直接合成するメル逆 LSP フィルタを提案し、有声音残差を容易に合成できることを明かにした。次に、メル逆 LSP のモデル化次数、量子化特性、及び補間特性について検討し、各々の最適化を図った。最後に、本システムを用いた合成音声の主観的な音質評価を行い、4.8kbits/sec 程度のビットレートで高音質な音声を合成できることを示し、本手法で提案したミクロな音源情報のモデル化システムを低ビットレートのボコーダへの応用が可能であることを明らかにした。

以上述べてきたように、音声生成モデルに基く LPC 分析合成システムにおいて、マクロな音源情報、即ちピッチ周期を正確に抽出・モデル化することが最重要な課題であり、第 2 章で提案した手法によりその目的を達成することができ、更に第 3 章において有声音残差からのミクロな音源情報抽出法を提案し、第 4 章でそれに基くミクロな音源情報のモデル化を提案することにより LPC 方式の限界性能の向上を実現し、高度情報化社会において要請されている低・中ビットレート音声符号化方式の実用化の可能性を明らかにした。これにより、本研究の目的が達成された。

謝　　辞

本研究は、著者が慶應義塾大学大学院理工学研究科博士課程在学中に、同大学理工学部教授森真作博士、小沢慎治博士の下で行ったものである。研究を遂行するに当たり、適切な御指導、御鞭撻を賜った森博士に深く感謝する。また、小沢博士には内容の詳細にわたり常に御指導御激励くださり、貴重な御助言、並びに適切な研究の方向を示していただいた。ここに深謝する次第である。

本論文の内容に関しては、同大学理工学部教授下郷太郎博士、同助教授中川正雄博士、東北大学応用情報学センター所長・教授城戸健一博士に貴重な示唆を賜った。こおに深く感謝する。

研究を進めるに際し、宮内新氏、恩田憲一氏、宮内ミナミ氏には、熱心かつ情熱的な御指導をいただいた。また、堀尾喜彦君、原崎秀信君、三石隆夫君、水野秀志君、全炳東君、荒井秀一君はじめ、小沢研究室及び森研究室の諸氏に御協力願った。ここに記して謝意を表する。

終わりに、研究を続けるに際し、影ながら深い愛情とともに援助と激励を下さった両親と兄夫婦に感謝する。そして、研究及び論文を完成するにあたり、心の支えとなり著者を励まし続けてくれた最愛の妻、恵子に心から礼を述べて結びとする。

参考文献

[1] C. E. Shannon, "A Mathematical Theory of Communication", Bell System Tech. J., Vol. 27, pp. 623-656, October(1968).

[2] 八塚陽太郎, " 中・高ビットレート音声符号化", 電子情報通信学会誌, Vol. 70, No. 4, pp. 392-400(1987).

[3] 宮川, 城戸 他共著, " ディジタル信号処理", 電子通信学会, pp. 43-45(1983).

[4] Dudley, H., "The vocoder", Bell Labs Record, 18, 4, pp. 122-126(1939).

[5] Kopec, G. E. Oppenheim, A. V. and Tribolet, J. M., "Speech analysis by homomorphic prediction", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-25, 2, pp. 40-49(1976).

[6] 深林, 鈴木, " 極ー零型の線形モデルによる音声分析", 信学論, J58-A, 5, pp. 270-277(昭 52).

[7] 石崎 俊, " 音声分析における極零モデルの次数の同定", 信学論, J60-A, 4, pp. 423-424(昭 52).

[8] 森川, 藤崎, "ARMA パラメータの同時推定法による音声分析", 音響学会音声研資, S76-53(昭 51).

[9] 嵯峨山, 古井, " 極零モデルによる音声スペクトルの最尤推定", 音響学会音声研資, S76-37(昭 52).

[10] Wiener, N., "Extrapolation, interpolation, and smoothing of stationary time series", MIT Press, Cambridge, Massachusetts(1966).

[11] Itakura, F. and Saito, S., "Analysis synthesis telephony based on the maximum likelihood method", Reports of the 6th Int. Cong. Acoust., C-5-5 (1968).

[12] Atal, B. S. and Schroeder, M. R., "Predictive coding of speech signals", Reports of 6th Int. Cong. Acoust., C-5-4(1968).

[13] Markel, J. D., "Digital inverse filtering-A new tool for formant trajectory estimation", IEEE Trans. Audio, Electroacoust., AU-20, 2, pp. 129-137(1972).

[14] J. Makhoul, "Linear Prediction-A Tutorial Review", Proc. IEEE, Vol. 63, pp. 561-580(1975).

[15] J. Makhoul, "Spectral Linear Prediction:Properties and Applications", IEEE Trans. Acoust. Speech, Signal Processing, ASSP-23, 3, pp. 283-296(1975).

[16] 三浦 種敏 監修, " 聴覚と音声", 電子通信学会, pp. 323-329(1980).

[17] Carr, P. B. and Trill, D., "Long-term larynx excitation spectra", J. Acoust. Soc. Am., 36, 11, pp. 2033-2040(1964).

[18] Rabiner, L. R. Cheng, M. J. Rosenberg, A. E. and McGonegal, C. A., "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-24, 5, pp. 399-418(1976).

[19] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. on Audio and Electroacoustics, AU-20, No. 5, pp. 367-377(1972).

[20] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor", IEEE Trans. Acoust. Speech and Signal Proc. ASSP-22, pp. 353-362(1974).

[21] M. M. Sondhi, "New Methods of Pitch Extraction", IEEE Trans. Audio and Electroacoustics, AU-16, No. 2, pp. 262-266(1968).

[22] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods

of Speech in the Time Domain", J. Acoust. Soc. Am., Vol. 46, No. 2, Pt. 2, pp. 442-448, August(1969).

[23] A. M. Noll, "Cepstrum Pitch Determination", J. Acoust. Soc. Am., 41, pp. 293-309(1967).

[24] 北脇, 板倉, 斉藤, "PARCOR 形音声分析合成系における最適符号構成", 電子通信学会論文誌, J61-A, 2, pp. 119-126(昭 53).

[25] 北脇, 板倉, "PARCOR 係数の非線形量子化と不均一標本化による音声の能率的符号化", 電子通信学会論文誌, J61-A, 6, pp. 543-550(昭 53).

[26] 菅村, 板倉, " 線スペクトル対 (LSP) 音声分析合成方式による音声情報圧縮", 電子通信学会論文誌, J64-A, 8, pp. 599-606(昭 56).

[27] B. Bogert, M. Healy, and J. Tukey, "The Quefrency Analysis of Time Series for Echoes", Proc. Symp. Time Series Analysis, Chap. 15, pp. 209-243(1963).

[28] 今井, 阿部, " 改良ケプストラム法によるスペクトル包絡の抽出", 電子通信学会論文誌, J62-A, 4, pp. 217-223(昭 54).

[29] 今井, 古市, " 高品質音声合成のためのインパルス列等価音源", J68-A, 11, pp. 1242-1252(昭 60).

[30] Atal, B. S. et al, "A new method of LPC excitation for producing natural-sounding speech at low bit rates", Proc. IEEE, ICASSP82, pp. 614-617(1982).

[31] Atal, B. S. and Schroeder, M. R., "Stochastic coding of speech at very low bit rate", Proc. ICASSP 84, pp. 1610-1613(1984).

[32] Un, C. K., Magill, D. T., "The residual-excited linear prediction vocoder with transmission rate below 9. 6 k bit/s", IEEE Trans., COM-23, 12, pp. 1466-1474(1975).

[33] 板倉文忠, " 新しい音声分析合成方式" ＰＡＲＣＯＲ", 日経エレクトロニクス, 2. 12, pp. 58-75(昭 48).

[34] 真野, 小沢, " ＬＰＣ分析を用いた残差波形の振幅包絡特性からの音源情報抽出", 電子通信学会論文誌, J67-A, 1, pp. 72-73(昭 59).

[35] Rabiner, L. R., Schafer, R. W. 著, 鈴木久喜 訳, " 音声のディジタル信号処理", コロナ社, p. 172, p. 199(昭 58).

[36] Brigham, E. Oran 著, 宮川, 今井 訳, " 高速フーリエ変換", 科学技術出版社 (昭 56).

[37] 安居院, 中嶋 著, " コンピュータ音声処理", 産報出版, pp. 132-134(1981).

[38] 板倉, 斉藤, " 統計的手法による音声スペクトル密度とホルマント周波数の推定", 電子通信学会論文誌, J53-A, 1, pp. 35-42(昭 45).

[39] 斉藤, 中田 著, " 音声情報処理の基礎", オーム社, P. 120(昭 56).

[40] 真野, 小沢, " ＬＰＣ有声音残差のピッチ同期分析に基く零点を有する励振パルスモデル", 電子情報通信学会論文誌 (A), J70-A, 6, pp. 960-968(昭 60).

[41] M. V. Mathews, J. E. Miller and E. E. David, Jr., "Pitch Synchronous Analysis of Voiced Sounds", J. Acoust. Soc. Am., 33, pp. 179-186(1961).

[42] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", J. Acoust. Soc. Am., 49, pp. 583-590(1971).

[43] Miller, R. L., "Nature of the vocal cord wave", J. Acoust. Soc. Am., 31, p. 667 (1959).

[44] 柏木, 中村, 高梨, " 線形予測残差のスペクトル包絡による話者識別", J68-A, 7, pp. 702-703(昭 60).

[45] 真野, 小沢, " ＬＰＣ有声音残差のピッチ同期メル逆ＬＳＰ分析合成方式", 電子情報通信学会

論文誌 (A), 掲載決定.

[46] S. S. Stevens, J. Volkman, ”A Scale for the Measurement of the Psycho-logical Magnitude Pitch”, J. Acoust. Soc. Am., Vol. 8, pp. 185-190(1937).

[47] Strube, H. W., ”Linear prediction on a warped frequency scale”, J. Acoust. Soc. Am., 68, 4, pp. 1071-1076(Oct. 1980).

[48] 今井, 住田, 古市, ” 音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ”, J66-A, 2, pp. 122-129(昭 58).

[49] 古井 著, ” ディジタル音声処理”, 東海大学出版会, P. 130(1985).

業績リスト

(査読付き論文)

1. ○真野淳, 小沢慎治：”ＬＰＣ有声音残差のピッチ同期メル逆ＬＳＰ分析合成方式”, 電子情報通信学会論文誌（A）,Vol. J71-A,3,pp.634-641,（昭 63）.
2. ○真野淳, 小沢慎治：”ＬＰＣ有声音残差のピッチ同期分析に基く零点を有する励振パルスモデル”, 電子情報通信学会論文誌,Vol.J70-A,6,pp.960-968,（昭 62）.
3. ○真野淳, 小沢慎治：”LPC 分析を用いた残差信号の振幅包絡特性からの音源情報抽出”, 電子通信学会論文誌（A）,Vol. J67-A,1,pp.72-73（昭 59）.
4. 宇佐聡史, 真野淳, 小沢慎治：”低域通過フィルタとダイナミックな値を用いた音声のピッチ抽出”, 電子通信学会論文誌（A）,Vol. J67-A,8,pp.857-858,（昭 59）.
5. Masayoshi.Nakamura, Atsushi.MANO, and Shinji.Ozawa : ”Coding of LPC Voiced Residual Using Pitch Synchronousanalysis and vector quantization”,The Journal of the Acoustical Society of Japan(E),(Submitted).

(研究会発表)

1. ○真野淳, 小沢慎治：”ＬＰＣ分析を用いた残差波形の振幅包絡特性からの音源情報抽出”, 情報理論とその応用シンポジウム,G-2, 第６回, 昭和 58 年 11 月

(国外学会発表)

1. ○Atsushi Mano, Shinji Ozawa :” Voiced Source Extraction from an Envelope of Residual Signals Using LPC analysis”,Acoustical Society of America, 108th Meeting,1984.
2. Satoshi Usa, Atsushi Mano, Shinji Ozawa :” A High Time-Resolution Pitch Detector”,Acoustical Society of America, 108th Meeting,1984
3. ○Atsushi Mano, Shinji Ozawa :”Suppression of Noise in Speech Using PARCOR Analysis”,Acoustical Society of America, 107th Meeting,1983.

(国内学会発表)

1. 中村, 真野, 小沢,：”残差を位相特性と雑音を用いて符号化するＬＰＣ分析合成”, 電子通信学会創立 70 周年記念総合全国大会講演論文 1334（昭 62）.
2. 中村, 真野, 小沢,：”一様でない振幅スペクトルを有するＬＰＣ残差の位相等化処理”, 電子通信学会創立 70 周年記念総合全国大会講演論文 1333（昭 62）.
3. ○真野, 荒井, 小沢,：”零点と位相特性に着目した有声音残差信号のピッチ同期分析”, 電子通信学会総合全国大会講演論文 1363（昭 61）.
4. 中村, 真野, 小沢,：”ベクトルの遷移辞書を用いた音声情報圧縮の一方式”, 電子通信学会総合全国大会講演論文 1592（昭 60）.
5. 宇佐, 真野, 小沢,：”低域通過フィルタとダイナミックな値を用いたピッチ抽出”, 電子通信学会総合全国大会講演論文 1653（昭 59）.
6. ○真野, 宮内, 小沢,：”ＰＡＲＣＯＲ分析における残差信号の振幅包絡特性からの音源情報抽出”, 電子通信学会総合全国大会講演論文 1471（昭 58）.
7. 新保, 真野, 小沢,：”時間的に変化する騒音を含む音声の明瞭度改善”, 電子通信学会総合全

国大会講演論文 64（昭 58）.
8. ○真野, 小沢, ：” ＰＡＲＣＯＲ分析から求めた時間域波形の振幅包絡特性を用いた適応量子化の一方式”, 電子通信学会情報・システム部門全国大会講演論文 5（昭 58）
9. ○真野, 小沢, ：” 騒音を含む音声の明瞭度改善 (2)”, 電子通信学会総合全国大会講演論文 80（昭 57）.
10. ○真野, 中井, 小沢, ：” 騒音を含む音声の明瞭度改善”, 電子通信学会情報・システム部門全国大会講演論文 245（昭 56）.