論文 / 著書情報
Article / Book Information

題目(和文)	基底膜モデルを用いた音声認識に関する研究
Title(English)	
著者(和文)	亀井宏行
Author(English)	HIROYUKI KAMEI
出典(和文)	学位:工学博士, 学位授与機関:東京工業大学, 報告番号:甲第1298号, 授与年月日:1981年3月26日, 学位の種別:課程博士, 審査員:
Citation(English)	Degree:Doctor of Engineering, Conferring organization: , Report number:甲第1298号, Conferred date:1981/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

基底膜モデルを用いた
音声認識に関する研究

昭和56年3月

指導教官　河原田弘　助教授

提出者　大学院 博士課程 電子システム専攻

亀井宏行

目　次

第1章　序論

音声の機械認識に関する研究は近年急速な発展を見，話者適応型の単語音声認識装置等すでに実用化されているものもある。しかしながら，不特定多数の話者を対象としたシステムの開発には，話者の個人性情報の取り扱い等困難な点が多く，いまだに芳しい成果は得られていない。また，連続音声になると，調音結合等の影響により，音響処理のみによる認識では音韻認識率は非常に低い値しか得られていない。ここでTable 1－1に，代表的な3手法による日本語の連続音声（単語音声）認識の実例を，母音認識を中心にまとめてみる。中津らによる新幹線座席予約システムでは[1]，母音標準パターンを話者別に作成するにもかかわらず，77、8%の母音認識率しか得られていない。三輪らによる都市名認識実験は[2]，不特定話者を対象にしているので50%以下の母音認識率しか得られていない。中島らによる声道形状のいくつかの特徴を用いた母音認識実験は[3]，連続音声中の母音に対して96%以上という非常に高い認識率を得ているが，閾値論理の数が4千数百（FORTRANで約8000 step）と膨大なもので実用化には難点がある。また，以上の実験で対象とされたのはいずれも成人男性ばかりであり，女性や子供の音声に対しても有効に動作するシステムを構成するまでには，さらに検討を加える必要があろう。

連続音声認識や単語音声認識では，音韻認識率の低さをいくつもの音韻候補を上げることによりカバーし，単語辞書や構文・意味論等の言語情報を用いて補なおうとする研究が行なわれているが[4],[5]，用途が限定され語い数が少ない場合ならともかく，語い数が増加した場合への拡張性は乏しい。従って，音声認識研究の究極の目的である不特定多数の話者を対象とし且つかなり広い用途に応用できる音声認識システムを開発するためには，音韻レベルでかなり高い認識率を得ることが重要なキーポイントとなると考えられる。特に，日本語は母音中心型の言語であり，又，子音の認識も後続母音があらかじめ正しく認識されていればかなり改善されることが期待できることから，母音に対して高

Table 1-1 連続音声認識の現状。

	最尤スペクトル分析	BPF群	声道形推定法
対象	新幹線座席予約 (文節ごと区切って発音)	都市名単語	和歌
語い数	112単語	166都市名	たいに歌(48音節)
母音認識法	LPC分析のαパラメータを標準パターンとし，入力の自己相関関数との尤度を求める。	Q=6，1/6 oct，29 ch (CF 250～6300Hz) のフィルタ群から得られるスペクトルのローカルピークを用いる。	声道形に現われる調音特徴を組み合わせた閾値論理網。
標準パターン	個人別に標準パターン作成。	成人男性5名により作成した標準パターンを，多少変更。	話者1名に対する閾値論理網から，順次1名ずつ対応させ閾値論理を拡大。(各音韻に複数のカテゴリ)
人数	成人男性 8名	成人男性 15名 (含．標準話者)	成人男性 9名
母音認識率	77．8％ (85．9％)	49％ (82％)	96％ いろは歌で母音認識システムを構成し，たいに歌を認識。
文節[単語]認識率	86％	83．2％	
研究機関	電々公社 中津，好田 (1)	東北大 (2) 三輪，新津，牧野，城戸	電総研 中島，鈴木，三国 (3)
備考	(　)内の数字は，複数の候補を上げた場合，その中に正しい母音が含まれる割合。		閾値論理の数，4千数百個 FORTRANで 約8000 step． 母音検出は人間が実行。

い認識率を得ることが，日本語音声認識システムを開発する上で重要である。

また，音声の機械認識が実用化されるに至り，実験室のような静かな環境下ばかりでなく，工場内のような高騒音下での使用に耐えうるものや，通信系と直接接続されて使用される場合には雑音や混信に強いものが，要求されるようになるであろう。人間は，多数の話者の会話音の中から，或いは他の背景雑音の中から特定の個人の声を聴き分け認識するという優れた能力を持っている。これは「カクテルパーティ効果」(6)としてよく知られている現象であるが，このカクテルパーティ効果の解明は，雑音や混信に耐えうる音声認識装置の開発に大いに役立つと考えられる。

本研究では，不特定話者対応連続音声認識システムを開発する上で特に重要な母音認識システムの開発と，人間の聴き分け機構（カクテルパーティ効果）を探ることを，大きな目的とする。

音声認識処理の手法としては，LPC分析法，ケプストラム法，声道関数推定法，バンドパスフィルタ群を用いる方法等，色々な方法があるが，(7)本研究では，聴き分け機構を探るという目的があるので，生体工学的立場から，聴覚系における最初の情報変換器である蝸牛基底膜のモデルを提案し，その基底膜モデルの出力を処理するという手法を採用する。生理学における聴覚系に関する研究の現状としては，基底膜振動に関してはかなり詳しく調べられている。(8),(9),(10),(11),(12)神経系に関しては，個々の神経の応答野や聴覚伝導路についてはかなり分かってきているが，(13)音声のような複雑なスペクトル構造を持つ入力に対して，それら神経系が総体としてどの様に機能するかは定かではない。実際の音声信号や合成音声を用いて神経応答を調べた例もあるが(14),(15)実験動物と人間の可聴域の違いがあり，又，脳の発達の程度の差により，ごく下位の神経系での比較しか出来ない。さらに生体工学の分野では聴神経系のシミュレーションもなされているが，(16)聴神経系での情報処理機構を解明するに足る知見は得られていない。基底膜モデルの出力から音声認識を行なおうという研究もいくつかなされているが，(17),(18),(19)認識に用いる特徴量や認識

手法は，従来からの音声波やそのスペクトルパターンに対する分析法とかわりのないものである。これは，聴覚系での特徴抽出機構や認識のメカニズムが全く解明されていない現状ではしかたのないものであり，本研究でも，基底膜モデルを用いる以外に，認識に関する手法については聴覚系の処理にこだわらない。

本論文の次章以下の構成は次のとおりである。第2章では，計算機処理に適した基底膜のディジタル回路モデルの構成について述べ，高速演算法についても検討している。第3章では，人間の聴き分け機構（カクテルパーティ効果）について検討を加え，聴覚系におけるパルス相関モデルを提案し，2名の話者により同時発声された持続母音から成る混合音声から，それぞれの成分母音のスペクトルを分離する方法について述べる。第4章では，成人男女，及び10～12才の子供達の発声した単母音の基底膜スペクトルパターンの観察から，不特定話者の単母音認識システムを提案する。本システムは，入力音声からピッチ抽出等を行ない話者の属性（年令・性別）を判断してから音韻の識別を実行するのではなく，話者の属性と音韻性をまとめて取り扱い，参照パターンとの距離をもとにした識別法と，スペクトルパターンの形態的特徴をもとにした識別法の2つの方法を組み合わせて認識処理を実行するという特徴を持ったものである。成人男性32名，同女性25名，10～12才の男子44名，同女子16名，計117名の発声した585個の単母音に対して98.3%という高い認識率を得ている。第5章では，第4章で示した単母音認識システムを応用して，不特定話者対応の連続音声認識システムを構成し，この母音認識システムが連続音声に対しても有効であることを検証する。子音に関しても，単に子音部と母音部を区別するだけにとどまらず，4つの大分類クラス，無声破裂音（p，t，k），無声摩擦音（s，sh，ch，ts，h），有声摩擦音（z），その他の有声子音（N，m，n，r，b，d，g，y，w）及び，補足的な4クラス，k，h，破擦音（ch，ts），撥音Nの各クラスに分類することを試みている。NHKニュースから録音した男性アナウンサー8名，女性アナウ

ンサー5名の連続音声データに対して，母音認識率が80.0%，子音の識別率が72.8%という値が得られたことを報告する。第6章では，本研究のまとめを行ない，残された問題についても言及する。

第2章　基底膜モデル

2.1 基底膜

人間の聴覚系では，鼓膜によってとらえられた音声振動は中耳の3個の耳小骨（ツチ骨 Malleus，キヌタ骨 Incus，アブミ骨 Stapes）を経て内耳の蝸牛（Cochlea）内のリンパ液に伝えられる。この振動は蝸牛内の基底膜（Fig. 2-1中の Cochlear partition に存在）に進行波を生ぜしめ，基底膜上の内・外有毛細胞により検出され高次中枢へと送られる。基底膜の振動形態については Békésy により詳しく調べられている。(8) それによると，基底膜は Fig. 2-2 に示されるように，音の周波数が低い場合は蝸牛頂に近い部分が大きく振動し，高い周波数の場合はアブミ骨底に近い部分が大きく振動する。そしてその共振位置は Fig. 2-3 のように周波数の対数とほぼ直線的な関係を示す。また，基底膜の最大振幅点の振幅とアブミ骨の体積変位との比は，Fig. 2-4 に示されるような周波数特性を持つ。つまり同一振幅入力に対しては，

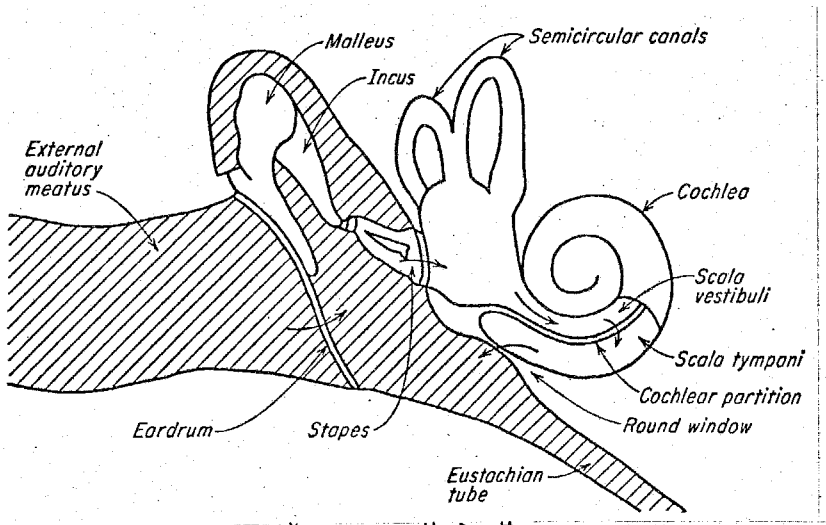


Fig. 2-1　耳の模式図。基底膜は Cochlear partition にある。（Helmholtz による，Békésy の著書より転載.）

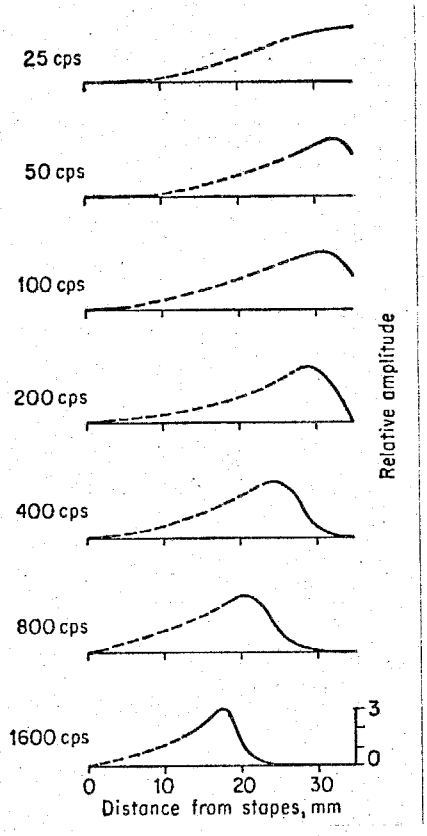


Fig. 2-2　基底膜の各周波数に対する共振パターン.（Békésy による.）

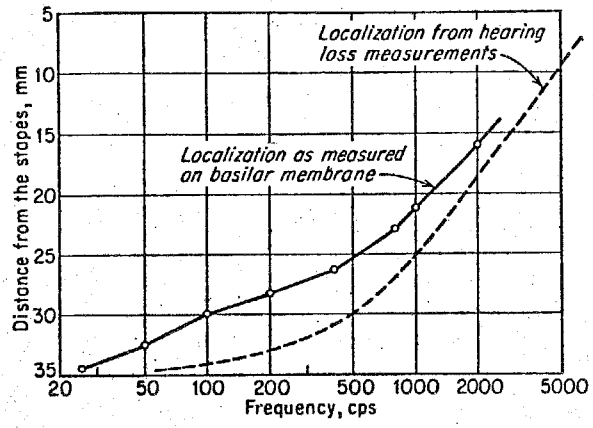


Fig. 2-3 各周波数に対し最大振幅を示す基底膜の位置. (Békésy による.)

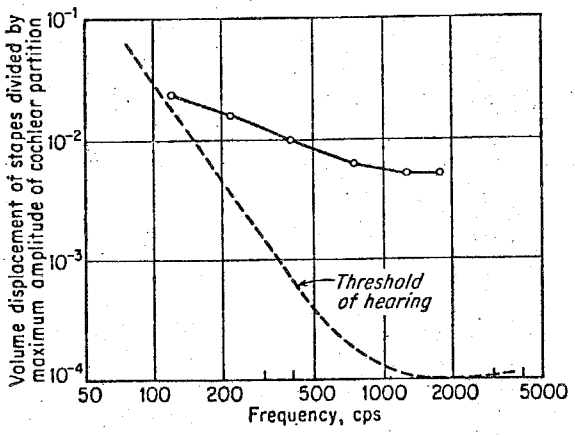


Fig. 2-4 アブミ骨の体積変位と基底膜の最大変位との比. (Békésy による)

基底膜の最大振幅は, 周波数が高くなるにつれ 約4~5 dB/oct の傾きで増加する。Fig. 2-5に基底膜の振幅・位相特性と, 基底膜上の各点の共振曲線を示すが, 共振のQ値は約1.7~2程度と低い。最近の研究によれば, 基底膜のQ値はさらに高く, また入力振幅の大小により非線形になるという報告もあるが, (9), (10), (11), (12) 測定方法の違いや対象生物の違いもあり, Békésy の結果との比較は困難である。そこで本研究では, Békésy の測定結果に基づいて基底膜モデルを設計する。

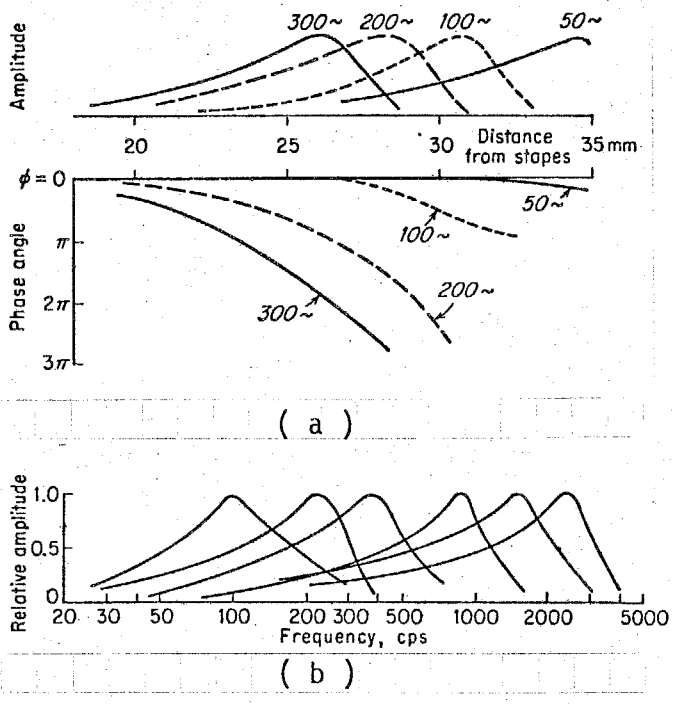


Fig. 2-5 (a) 4つの低い純音に対する基底膜の振幅・位相特性. (振幅は最大値を規格化) (b) 基底膜上の6点の共振曲線. (いずれも Békésy による.)

2.2 基底膜モデル

基底膜振動の定式化はFlanagan[20]によって成され，分布定数回路や梯形回路の形でシミュレーションも数多く成されている。[21],[22],[23],[24],[25]しかしそれらのモデルはアナログ回路や数学モデルであり，音声認識等の目的のために各種音声信号に対する基底膜振動の様子を色々な角度から詳しく観察したり計算機処理するには不向きである。そこで計算機処理に適した構造の新しい基底膜モデルを開発する必要がある。筆者は，全く構造の等しい2次のディジタルフィルタを54段縦続接続するという簡単な回路で基底膜モデルを構成することに成功した。以下本節でその基底膜モデルについて説明する。

2.2.1 単位フィルタ

基底膜はその周波数特性から，低域通過フィルタをそのカット・オフ周波数の高いものから順次縦続接続してシミュレートできることがわかる。そこで，本研究では，低域通過フィルタとしてFig.2-6に示される2次のディジタルフィルタを採用する。k段目の単位フィルタの入力は，前段のフィルタの出力 $y_{k-1}(n)$（そのZ変換を $Y_{k-1}(z)$ とする）で，出力 $y_k(n)$（$Y_k(z)$）は後段のフィルタの入力となる。また遅延レジスタ間の差分 $\dot{y}_k(n)$（$\dot{Y}_k(z)$）は基底膜の第kチャネル出力を与える。この単位フィルタの伝達関数 $H_k(z)$ は（2-1）式で与えられ，その振幅特性 $|H_k(e^{j\omega T})|$（T：サンプリング間隔）はFig.2-7のように設定する。

$$H_k(z)=\frac{Y_k(z)}{Y_{k-1}(z)}=\frac{a_3^k}{1-a_1^k z^{-1}-a_2^k z^{-2}} \qquad (2-1)$$

このフィルタの係数（a_1^k, a_2^k, a_3^k）は，（2-2）式の3条件から得られる連立方程式を解くことにより求められる。（解法は付録参照）

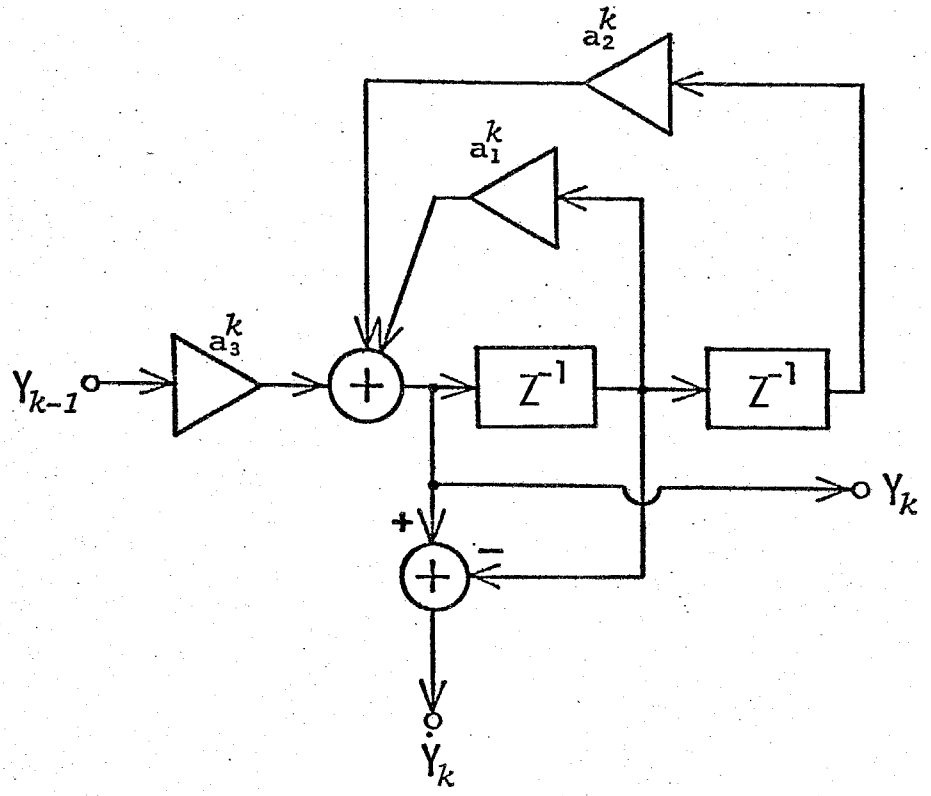


Fig. 2-6　基底膜モデルを構成する単位フィルタ.

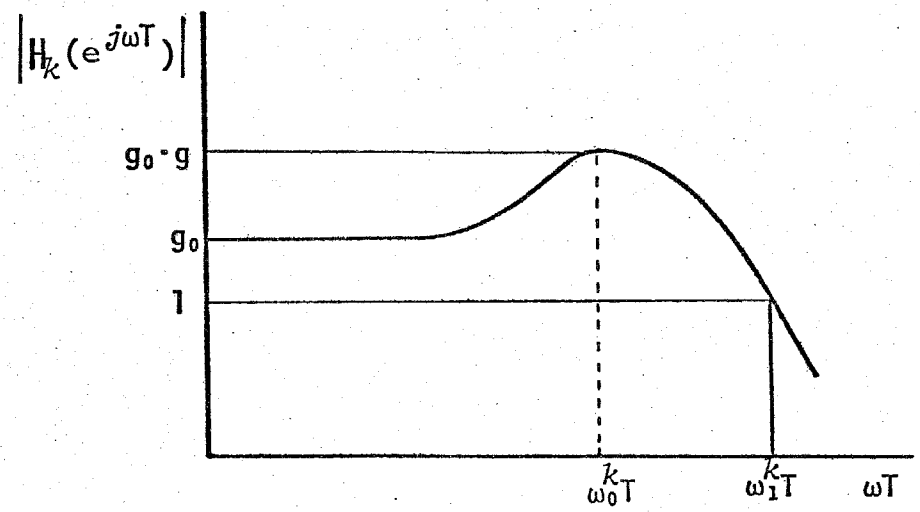


Fig. 2-7　単位フィルタの周波数特性 ($H_k(z) = Y_k(z) / Y_{k-1}(z)$).

$$\left.\begin{aligned} &|H_k(1)| = g_0 \quad (\text{at } \omega = 0) \\ &|H_k(e^{j\omega_0^k T})| = g_0 g \\ &\frac{d}{d(\omega T)} |H(e^{j\omega T})| \Big|_{\omega = \omega_0^k} = 0 \end{aligned}\right\} \quad (2-2)$$

その結果を以下に示す。

$$a_1^k = 2P - \sqrt{4P^2 - A} \quad (2-3)$$

$$a_2^k = \frac{a_1^k}{a_1^k - 4P} \quad (2-4)$$

$$a_3^k = g_0 (1 - a_1^k - a_2^k) \quad (2-5)$$

ここで

$$A = Q - \sqrt{Q^2 - 16P^2} \quad (2-6)$$

$$Q = \frac{2P^2 g^2 + 2g^2 - 4P}{g^2 - 1} \quad (2-7)$$

$$P = \cos\omega_0^k T = \frac{(g^2-1)\cos\omega_1^k T + \sqrt{(g^2-1)(g_0^2 g^2 - 1)}}{g^2 - 1 + \sqrt{(g^2-1)(g_0^2 g^2 - 1)}} \quad (2-8)$$

単位フィルタの振幅特性を与えるパラメータ g_0, g, ω_1^k およびサンプリング間隔 T を決定すれば a_1^k, a_2^k, a_3^k は (2-3) 式～(2-8) 式により算出される。

2.2.2 全体構成

基底膜モデルは，前項の単位フィルタ全54段の縦続接続で構成する。基底膜モデルは54個の出力端子を持つが，この出力端子を「チャネル」と呼ぶ。音声入力 $x(n)$ $(X(z))$ から 第kチャネルの基底膜変位出力 $y_k(n)$ $(Y_k(z))$ に至る伝達関数 $G_k(z)$ は（2－9）式で与えられる。

$$G_k(z) = \frac{Y_k(z)}{X(z)} = (1-z^{-1}) \prod_{n=1}^{k} H_n(z) \qquad (2-9)$$

各単位フィルタの f_1^k（$f_1^k = \omega_1^k / 2\pi$，振幅特性が1となる周波数）は16 kHzから1/6オクターブずつ減少するように（2－10）式で与える。

$$f_1^k = f_m \times 2^{-(k/N)} \qquad k = 1 \sim 54 \qquad (2-10)$$

$f_m = 16000\ Hz$，$N = 6$（1オクターブ当りの段数）

よって基底膜モデルは1オクターブ当り6段で，16 kHzから約30 Hzの9オクターブの周波数に対して振動する。振幅特性を決定する他のパラメータ g_0，g については，モデルの構成上全段一定とした方が望ましいので，この条件下で基底膜モデルの振幅特性が実際の基底膜の特性に近づくように試行実験を重ね，$g_0 = 1.06$，$g = 1.30$ と決定した。なお，音声信号のサンプリング周波数は 40 kHz（サンプリング間隔 $T = 25\ \mu sec$）である。

9オクターブ全域にわたり，サンプリング間隔一定のまま，各単位フィルタの係数を計算すると，最終段の単位フィルタの係数と初段フィルタの係数では10倍以上（$a_1^{54}/a_1' \simeq 13$）の違いが出る。また最終段近くになると，係数は一定の値に収束してしまう（$\lim_{k\to\infty} a_1^k = 2$）。このことは，ディジタル演算上，けた落ち，丸め誤差を生じ，回路構成上レジスタ長を長くとらねばならなくなるので好ましくない。ディジタルフィルタでは，周波数特性は ωT の関数であるので，$\omega T = (\omega/2) \times 2T$ の関係から サンプリング間隔を2倍にすると周波数特性が1オクターブ低周波側へシフトしたフィルタが得られる。この性質を利用し，基底膜モデルでは最初の2オクターブ分12段の

の単位フィルタの係数のみ計算し，3オクターブ目以後のフィルタ群は2オクターブ目のフィルタ群と同じ係数を与え，そのかわりに3オクターブ目以後は1オクターブ（6段）ごとにサンプリング周波数を1／2ずつ下げていく。この操作により，演算回数はサンプリング間隔一定とした場合に比べ3分の1となり，高速化もはかれる。また，基底膜出力 $y_k(n)$ は差分として与えられるので，各オクターブ群の初段フィルタ（13，19，25，31，37，43，49段目のフィルタ）への入力を1／2倍することにより，サンプリング間隔の違いにより生ずる $y_k(n)$ の振幅のずれを補正する。Fig. 2-8に，基底膜モデルの全体構成を示し，Fig. 2-9に基底膜モデルの周波数特性を示す。

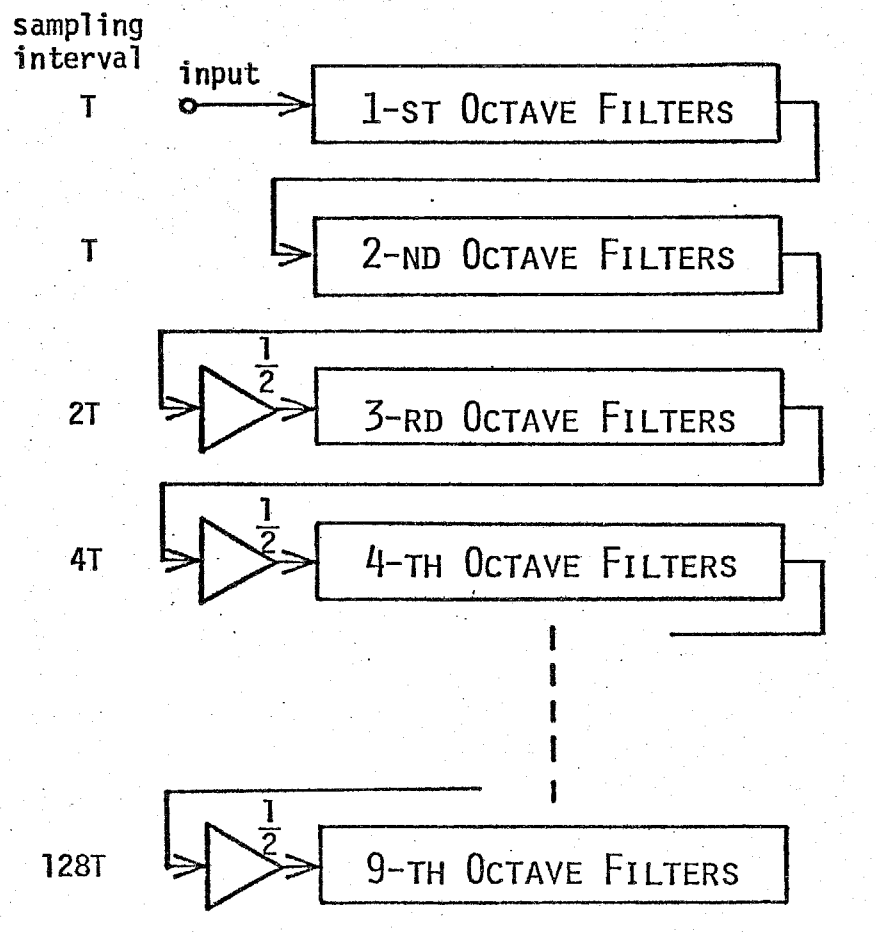


Fig. 2-8 基底膜モデルの全体構成図.

基底膜モデルのQ値は約2となり，また最大振幅値も約5 dB/octで周波数の上昇に伴ない増加する傾向を示し，実際の基底膜をよく近似している。また，第kチャネルの共振周波数（以下，特徴周波数 CFと略す）は，ほぼ f_k となる。Table 2-1に，単位フィルタの係数1～12段分を示す。

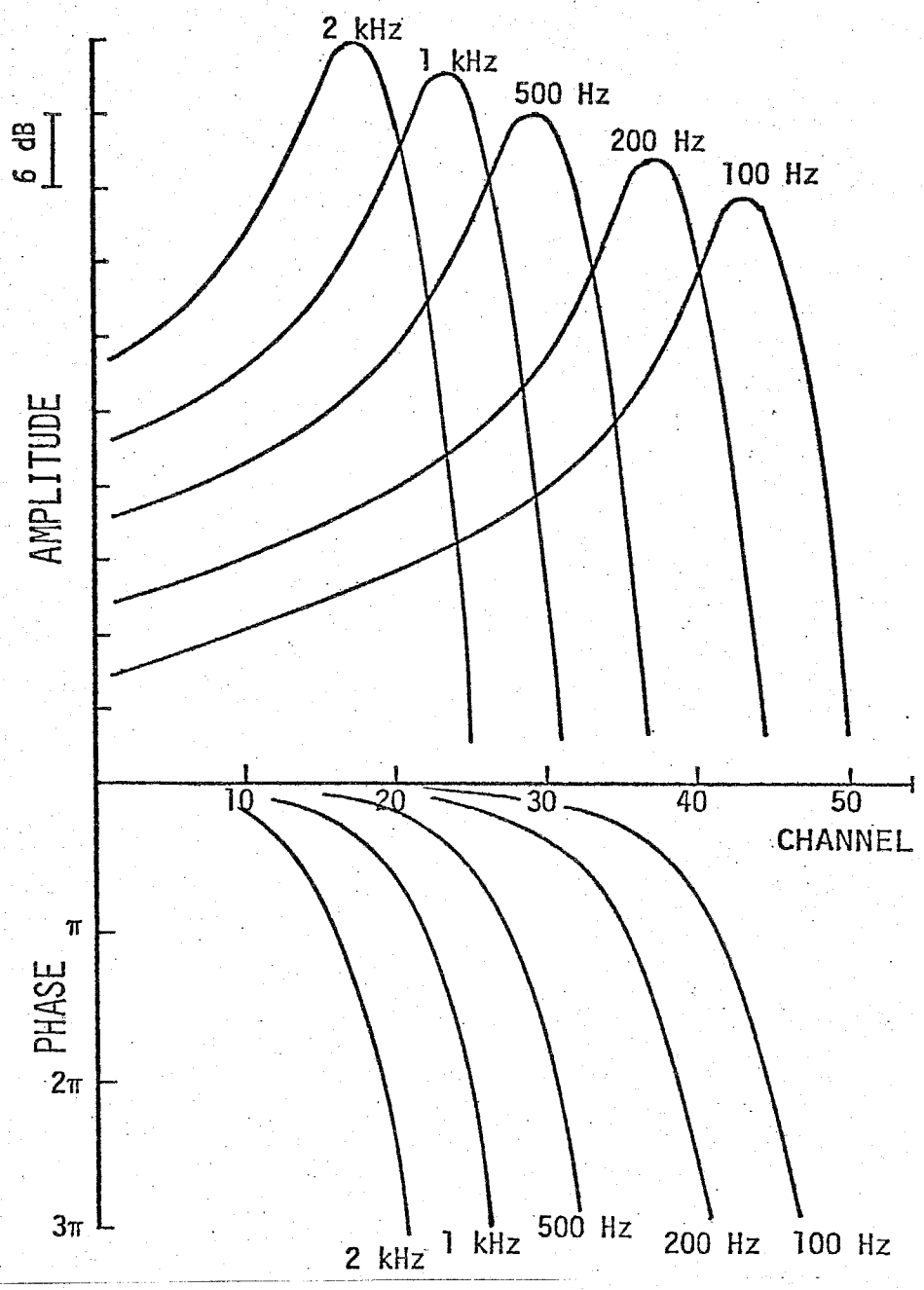


Fig. 2-9　基底膜モデルの周波数特性.

Table 2-1 基底膜モデルを構成する単位フィルタの係数.

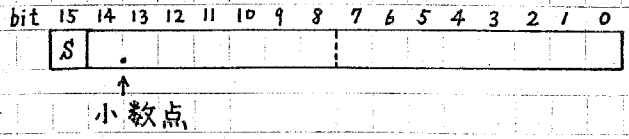
k	a_1^k	a_2^k	a_3^k
1	0.1563116	-0.1910259	1.0967970
2	0.2486917	-0.2230688	1.0328398
3	0.3615693	-0.2605466	0.9529159
4	0.4888072	-0.3017433	0.8617123
5	0.6234566	-0.3450841	0.7649251
6	0.7591089	-0.3892543	0.6679541
7	0.8907087	-0.4332350	0.5750779
8	1.0147622	-0.4762724	0.4892008
9	1.1291895	-0.5178269	0.4119557
10	1.2330370	-0.5575340	0.3439668
11	1.3261421	-0.5951526	0.2851512
12	1.4088616	-0.6305355	0.2349743

2.3 基底膜演算法

基底膜モデルの演算を計算機内で実行する場合，あるいはそのハードウェアを構成する場合，浮動小数点型データの乗算は非常に時間を要する。この節では，基底膜の演算を整数型（固定小数点型）として実行するための手法を示す。本研究で使用した計算機（HP2100A，HP2113E）に即して説明する。データ長は符号つき16ビットとし，負数は2の補数で表わす。

(1) 係数の符号化

Table 2-1を見ると，各係数の絶対値 $|a_i^k| < 2$ であることがわかる。そこで小数点の位置を bit 14 と bit 13 の間にとればよい。



例).

$a_1^{12} = 1.4088616$ → 01.011010:00101010

8進数表示 055052B

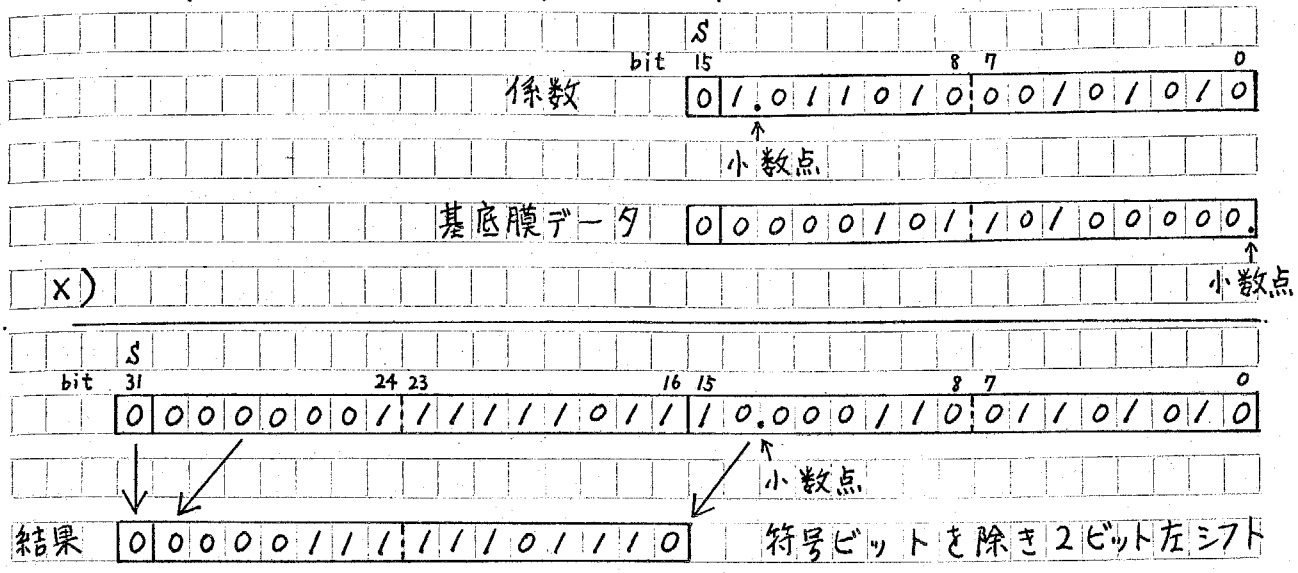
Table 2-2に8進数で表現した係数を示す。

Table 2-2 8進数（16ビット）で表現した単位フィルタの係数.（負数は2の補数で表示。）

k	a_1^k	a_2^k	a_2^k
1	005001B	171706B	043061B
2	007752B	170671B	041032B
3	013443B	167523B	036374B
4	017510B	166260B	033446B
5	023746B	164752B	030364B
6	030225B	163426B	025277B
7	034401B	162105B	022316B
8	040361B	160604B	017517B
9	044104B	157333B	015135B
10	047352B	156121B	013003B
11	052337B	154751B	011077B
12	055052B	153645B	007411B

（2） 乗算

係数は固定小数点型データなのに対し，音声データや基底膜変位は整数として取り扱う。係数とそれらデータの乗算を実行すると，結果は32ビットになり，小数点は下位16ビットのbit14とbit13の間にくる。そこで，32ビット全体を符号ビットを除き2ビット左シフトして，上位16ビットの整数部のみ取り出すという操作を行なう。以下，この手順を図解する。



この操作をアセンブラ言語（HP-Assembler）で書くと以下のようになる。

LDA CE --- CEの内容をAレジスタへ格納。

MPY BM --- Aレジスタの内容にBMの内容を乗じ，結果の上位16ビットをBレジスタに，下位16ビットをAレジスタに格納する。

ASL 2 --- BレジスタとAレジスタの内容を合わせて2ビット左へシフトする。（符号ビットはそのまま.）

STB NBM --- Bレジスタの内容を NBMに格納。

ここに CE：係数データのメモリ番地

BM：基底膜データのメモリ番地

NBM：演算結果を格納するメモリ番地

(3) オーバーフロー問題

以上の演算を実行する際に問題となるのは，オーバーフローの問題である。今，k段目の単位フィルタの出力 $y_k(n)$ ($Y_k(z)$) について考えてみる。音声入力を $x(n)$ ($X(z)$) とすると，$x(n)$ から $y_k(n)$ への伝達関数 $G'_k(z)$ ($= Y_k(z)/X(z)$) は，(2-11) 式で与えられる。

$$G'_k(z) = \prod_{n=1}^{k} H_n(z) \qquad (2-11)$$

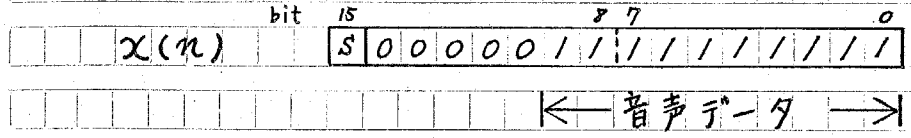
各周波数に対し $|G'_k(e^{j\omega T})|$ を計算すると，その形は Fig. 2-9 とほぼ同じで，ただ各周波数に対する最大振幅値が Fig. 2-9 とは逆に約 -7dB/oct の傾きとなる。つまり周波数が低くなる程最大ゲインは大きくなることを表わしている。実際の音声信号は 100 Hz 以下の成分は持たないことから，今，100 Hz の純音について考えてみる。100 Hz の純音が入力した場合，$y_k(n)$ が最大値を示すフィルタは 43 段目で，その時 $|G'_{43}(e^{j\omega T})|_{f=100Hz} = 103.0$ である。また，42 段目は $|G'_{42}(e^{j\omega T})|_{f=100Hz} = 88.5$ である。しかしながら基底膜モデルでは，Fig. 2-8 で示したように 1 オクターブ (単位フィルタ 6 段) ごとに，1/2 の減衰が与えられているので，43 段目に至るまでに 1/64，42 段目に至るまでに 1/32 倍される。よってこの減衰器の効果により実効的なゲインは，

$$|G'_{43}(e^{j\omega T})|/64 = 1.609 \quad , \quad |G'_{42}(e^{j\omega T})|/32 = 2.765 \quad \text{at } 100\,Hz$$

となり，$x(n)$ としては最低 2 ビット (4 倍) の余裕が必要であることがわかる。さらに係数との乗算で，係数の絶対値 $|a_i^k|$ は 2 未満であるので，1 ビットの余裕が必要であり，加算 $y_k(n) = a_1^k y_k(n-1) + a_2^k y_k(n-2) + a_3^k y_{k-1}(n)$ の演算で最大で見積っても 2 ビットの余裕が必要である。以上まとめると，音声入力としては合計 5 ビットの余裕があればよいことになる。

$x(n) \longrightarrow y_k(n)$	最大4倍	2ビット
$a_i^k y_k(n-i)$ $(a_3^k y_{k-1}(n))$	最大2倍	1ビット
$y_k(n) = \sum_{i=1}^{2} a_i^k y_k(n-i) + a_3^k y_{k-1}(n)$	最大3倍	2ビット
	合計	5ビット

つまり，最初に与える音声データを，16ビット中下位10ビット以内に格納すれば，絶対にオーバーフローは生じない。



次章以後の実験では，音声信号は40KHzのサンプリング周波数で，符号つき10ビットのディジタルデータにA/D変換され，ミニコンピュータ内にプログラムされた基底膜モデルに入力される。その際，符号を除いた9ビットの音声データは，16ビット中 bit 9 から bit 1（符号は bit 15）にセットした。

以上の基底膜演算のアルゴリズムは，ハードウェア作製の際にも応用され有効性が証明されている。(26),(27)

さて，この演算法を用いて基底膜モデルをハードウェアで構成した場合，実時間処理が可能であるかどうかを考察してみよう。1つのサンプリングされた音声データに対して，演算が実行される単位フィルタの段数は各時刻まちまちである。最終9オクターブ目のフィルタ群（48～54段目のフィルタ）は，128個のデータにつき1回の演算しか実行しないので，今サンプリングされたデータ128個（時間にして $128 \times 25\,\mu sec = 3.2\,msec$）について演算を終了するに必要な時間について検討してみよう。基底膜演算で最も時間を要するのは，係数との乗算である。TTL-ICを用いて16ビット×16ビ

ットの乗算器を製作すると，例えば，テキサス・インスツルメンツ社製のLSI SN74S274（4ビット×4ビット乗算器）を16個使い構成すると，132 nsecで乗算が完了する乗算器が構成できる。[33] 基底膜モデルのハードウェア構成として，単位フィルタ1台のみハードウェアで構成し，係数レジスタの値を書き換えることにより全54段分の機能を果すというシステムを考え，乗算器は3台用意すると仮定する。そうすると，単位フィルタ1段の演算で乗算に要する時間は132 nsecで済む。128個の音声データを演算するのに，演算を実行する単位フィルタの総数は，2298個であるので，全乗算に要する時間は，0.3 msecである。加減算に関しても同様に見積ると，総計0.3 msec程度で完了できる。3.2 msecの20%未満の時間で四則演算が終了できるので，3.2 msec以内で全演算を終了することは可能である。そこで，128 word（1 word 16ビット）の音声データ用のメモリを2つ用意しておき，1方が音声データを取り込んでいる間，他方のメモリの音声データについて基底膜の演算を実行するというシステムを考えれば，実時間処理が可能となる。

第3章　混合音声の聴き分け機構

3.1 まえがき

人間は，多くの話者の会話音の混在する中から特定の個人の音声に注目してその個人の話を聞きとることができる。これは，「カクテル・パーティ効果」としてよく知られている現象であり，この聴き分け機構の解明は，人間の情報処理機構を探る上で重要な意味を持つ。また，工学的応用の面からは，雑音に強い音声認識装置を開発する上に役立つものと考えられ，純粋に工学的立場からのカクテル・パーティ効果への2~3のアプローチ例はあるが，[28],[29] 実際の人間の情報処理機構にせまるものはまだ無い。現在，実用化段階に入った音声認識装置は，今後多方面にわたって利用されるようになることが確実であるが，それに伴い使用環境も多様化し，工場等の高雑音下での使用に耐えうるものが要求されてくるであろう。雑音中からの音声信号の検出・分離という問題の解決に，この聴き分け機構の解明は1つのヒントを与えるであろう。

音声情報中のどの個人性情報に着目して人間が聴き分けを行なっているかは定かではない。しかし，よくハーモニーのとれた合唱の中から特定の個人の声を聴き分けるのは非常に難しいという経験等から，主にピッチの違いに着目して聴き分けを行なっているものと考えられる。ピッチの異なる2つの成分からなる信号から各々の成分信号を分離するには，2つの成分信号間に相関がなければ，各々の成分信号のピッチに同期して同期加算を行なえばよいが，この場合各成分のピッチがあらかじめ分っていなければならない。しかしながら，音声の認識には信号波形そのものを再生する必要はなく，そのスペクトル包絡が分離されればよい。基底膜は，バンドパス・フィルタ群と考えられるので，基底膜各位置ごとの各成分信号の平均振幅が分離できれば，各々の成分信号のスペクトル・パターンが得られる。基底膜各位置ごとの各成分信号の平均振幅は，その位置の振動波形の自己相関関数を求めることにより，各ピッチに対応して現われるピークからピッチ周期（T）を求め，時間差Tにおける相関関数の値

から求めることができる。

聴覚神経系に相関関数を求める機能があることは，従来から漠然と指摘されていて，ピッチ検出のシミュレーション等試みられている。[30] 本章では，前章で提案した基底膜モデルを用い，基底膜各位置出力から各成分音のピッチ検出とスペクトルパターン（基底膜振幅パターン）の分離を行なう聴覚系モデルを提案し，2名の話者により同時発声された持続母音からなる混合母音を用いて行なった分離実験について報告する。

3.2 聴覚系モデル

神経系における相関機構を仮定して，次の手順で聴き分け機構をシミュレーションした。Fig. 3-1に，この手順のブロック図を示す。

(1). 音声信号は40kHzのサンプリング周波数でA/D変換され，計算機内にプログラムされた基底膜モデルに入力される。

(2). 基底膜の各チャネル出力を，有毛細胞*の整流効果を考慮して半波整流する。

(3). 1次ニューロンでの応答を模擬するために，(2)で得られた半波整流波形からピークを検出し，そのピーク位置に対応して幅200μsec，高さはそのピークの大きさのパルスを発生する。パルス幅の決定は生理学的事実に基づくものではなく，このシステムが離散系であるので，サンプリング時刻がピークに合致しない場合にその影響を少なくするために十分な幅を持たせた。また，実際の神経系では，振幅情報はパルス頻度という形で伝送されるが，このモデルでは単純化のためパルス高で代用する。

* 基底膜上に存在し，基底膜振動を検出し神経系へ伝達する機能を有する細胞。

(4). 得られたパルス列の自己共分散関数を求める。ここで自己相関関数でなく，自己共分散関数を用いるのは，以下の理由による。

今，2つの周期信号 $g(t)$（周期 T_1），$f(t)$（周期 T_2）の和信号 $x(t)$ を考える。

$$x(t) = g(t) + f(t) \qquad (3-1)$$

$x(t)$ の自己共分散関数 $R_{xx}(\tau)$ は（3－2）式で与えられる。

$$R_{xx}(\tau) = R_{gg}(\tau) + R_{ff}(\tau) + R_{gf}(\tau) + R_{fg}(\tau) \qquad (3-2)$$

ここで $R_{gg}(\tau) = E[(g(t)-\bar{g})(g(t+\tau)-\bar{g})]$

$R_{ff}(\tau) = E[(f(t)-\bar{f})(f(t+\tau)-\bar{f})]$

$R_{gf}(\tau) = E[(g(t)-\bar{g})(f(t+\tau)-\bar{f})]$

$R_{fg}(\tau) = E[(f(t)-\bar{f})(g(t+\tau)-\bar{g})]$

$\bar{g} = E[g(t)]$ ，$\bar{f} = E[f(t)]$

$g(t)$，$f(t)$ の周期性を考慮し，$R_{xx}(T_1)$ を求めると（3－3）式となる。

$$R_{xx}(T_1) = R_{gg}(0) + R_{ff}(T_1) + R_{gf}(T_1) + R_{fg}(0) \qquad (3-3)$$

（3－3）式で，$g(t)$ と $f(t)$ の相関が無ければ（3－4）式の近似が成立し，また $T_1 \neq T_2$ から $R_{ff}(T_1) \simeq 0$ となり（3－5）式が得られる。

$$R_{gf}(T_1) \simeq 0 \ , \quad R_{fg}(0) \simeq 0 \qquad (3-4)$$

$$\therefore \quad R_{xx}(T_1) \simeq R_{gg}(0) \qquad (3-5)$$

（3－5）式は，$R_{xx}(T_1)$ を求めれば $g(t)$ の平均振幅が求められることを示している。（2），（3）の手順で得られるパルス列は非負値関数であるので，自己相関関数を用いると，$R_{gf}(T_1)$，$R_{fg}(0)$，$R_{ff}(T_1)$ に相当する項が0に近似できず，又，それらの値がすべてのチャネルにわたり一定にはならないので，スペクトルはうまく分離できない。

(2)～(4)の操作を基底膜各チャネル出力に対して行なう。また，(3)の操作でパルス幅を200 μsec としたが，19チャネル(CF 2 kHz)以前では，パルスが重複する可能性があるので，(3)の操作を省略し直接(4)の操作を行なうものとする。

(5). 各チャネルの自己共分散関数(以下 R(τ; k)(k: チャネル番号)と表記する)から，成分音声のピッチ周期 Ti を検出し，その時間差 τ = Ti における各チャネルの値 R(Ti; k) を見ることにより スペクトルパターンを分離する。ピッチ検出の仕方は，次節で実例を示しながら説明する。

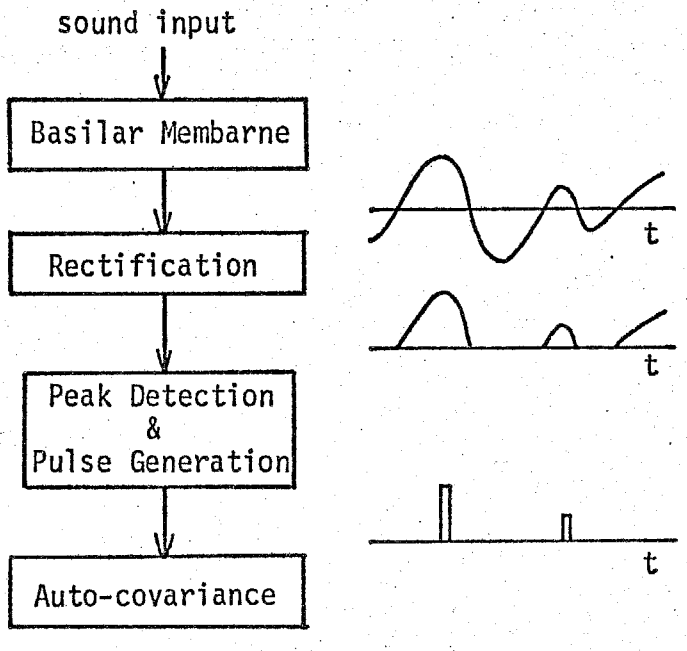


Fig. 3-1 聴覚系における神経パルス相関モデル.

3.3 混合母音の分離実験

2名の話者により同時発声された持続母音よりなる混合音声，あるいは別々に発声し録音された持続母音をミクサで合成した混合音声について，そのスペクトルパターンの分離実験を行なった。母音の特徴は，13チャネル（CF 4 kHz）から42チャネル（CF 125 Hz）の範囲にあることが明らかであるので，この帯域の各チャネルに対して自己共分散関数 $R(\tau; k)$ を求めた。また，時間差 τ については，母音のピッチが100 Hz 以上であることから，$\tau = 0$ から 10 msec まで，0.1 msec きざみで計算した。1つの共分散関数を求めるに用いたデータ（基底膜出力）長は，51.2 msec である。この場合，共分散関数は，FFTを用いて計算するよりも定義式に従い直接計算する方が速いので，直接法により計算した。

以下，各種母音の組み合わせによる成分母音のピッチ検出ならびにスペクトルパターン分離抽出の結果を示す。ここで用いられた混合音声はすべて試聴の結果，聴き分けができたものばかりである。

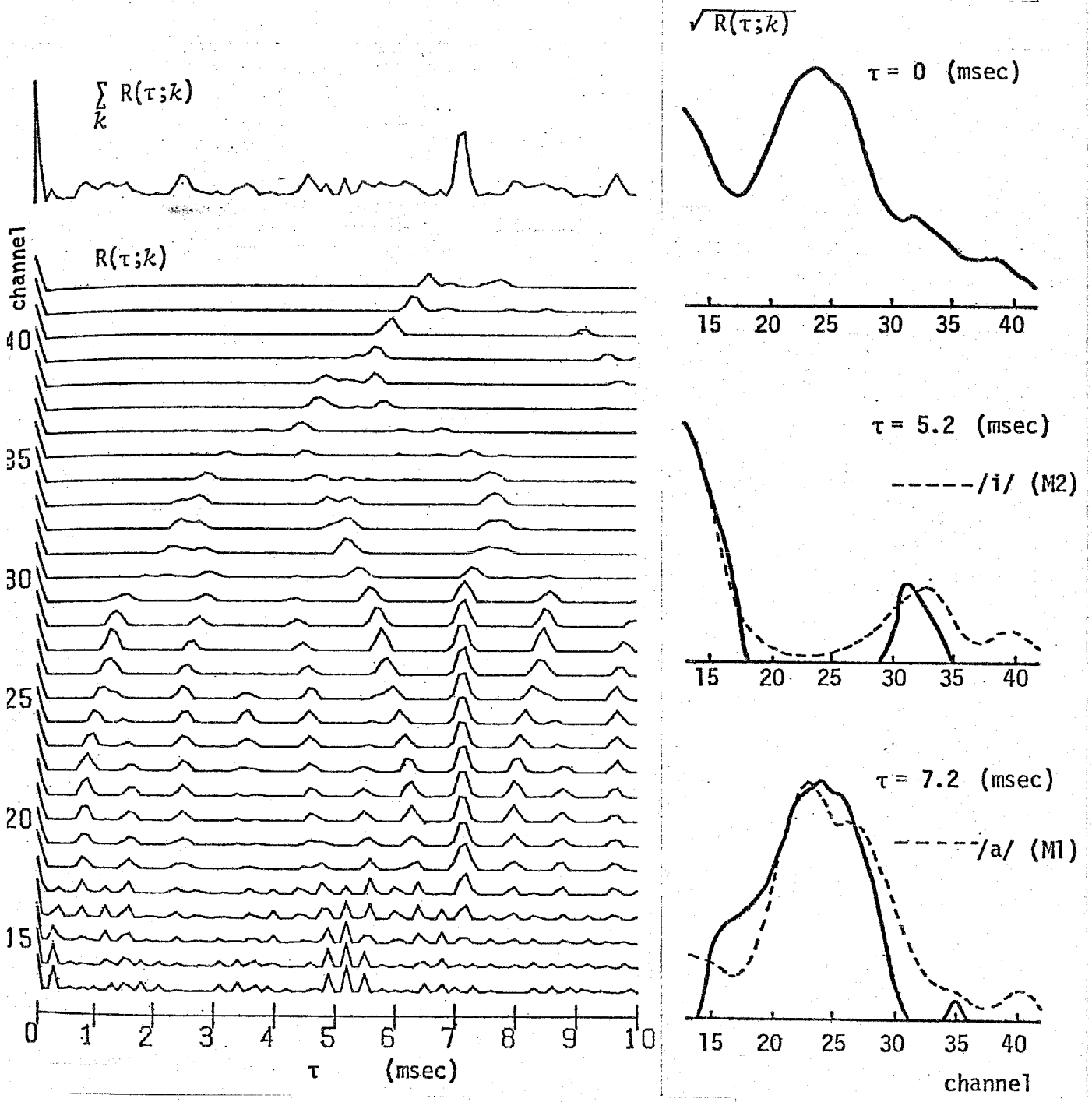
(1). 成分母音及びピッチが異なる場合.

ここではまず，成分母音の種類が異なり，且つ一方のピッチが他方のピッチに重なりもせず，またその整数倍にもなっていない場合についての結果を報告する。

一例を Fig. 3-2 に示すが，入力音声は男性（M1）の発声による /a/ と，別の男性（M2）の発声による /i/ の混合音声である。Fig. 3-2 (a) は，自己共分散関数 $R(\tau; k)$ を各チャネルごとに $R(0; k)$ で規格化して表示したものである。成分母音のピッチは，周期関数の自己共分散関数がその周期ごとにピークを持ち，そのピークに対して対称な形を示すという性質を利用し，同図 (a) から検出する。つまり，幾つかのチャネルにわたり同じ位置（時間差）にピークを持ち，更にそのピークに対して対称な形を示しているパターンを見つけ出し，そのピーク位置の τ を読み取る。

Fig. 3－2（a）からは，13～18チャネルのτ＝5．2 msecのピークと，15～29チャネルのτ＝7．2 msecのピークが5，T_1＝5．2 msec，T_2＝7．2 msecと2成分母音のピッチが検出できる。Fig. 3－2（b）は，τ＝0，τ＝5．2 msec，τ＝7．2 msec位置での自己共分散関数の値の平方根，つまり$\sqrt{R(0;k)}$，$\sqrt{R(T_1;k)}$，$\sqrt{R(T_2;k)}$を各チャネル方向にとって図示したものである。また，同図中破線のパターンは，それぞれの成分母音を単独で入力した場合の基底膜平均振幅パターン（スペクトルパターン）で，$\sqrt{R(T_i;k)}$（$i=1,2$）とは最大値を一致させて表示してある。$\sqrt{R(0;k)}$の図では，個々の成分母音のスペクトルパターンを加え合わせた形を示しているが，$\sqrt{R(T_1;k)}$，$\sqrt{R(T_2;k)}$の図では，それぞれ破線のパターンとよく似ていることから，各成分母音のスペクトルパターンが分離されていることがわかる。

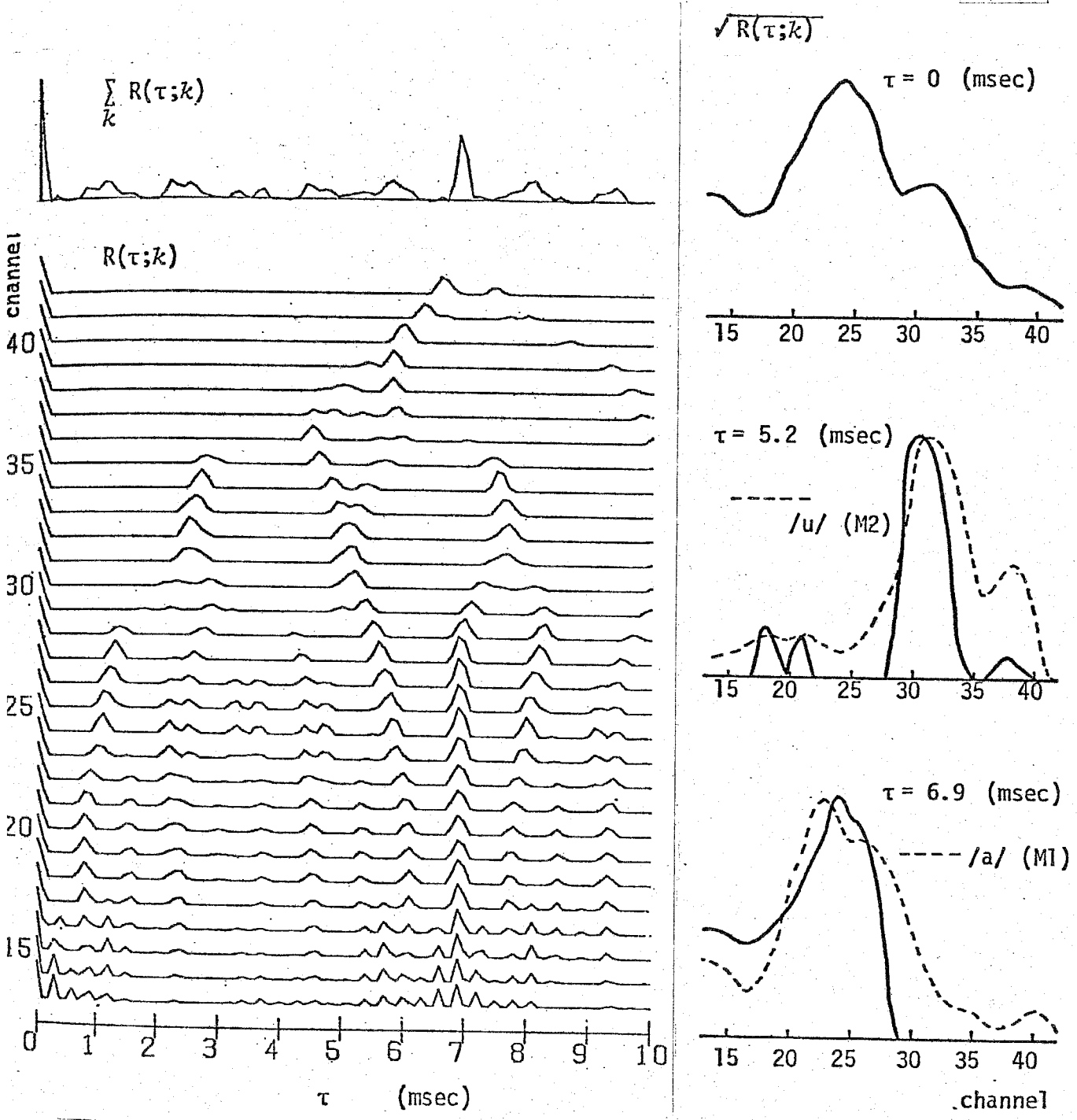
Fig. 3－3には，Fig. 3－2と同一話者で母音の組み合わせが異なる場合（男性（M1）の/a/＋男性（M2）の/u/）の結果を示すが，T_1＝5．2 msec，T_2＝6．9 msecとピッチが検出でき，また，分離スペクトルパターンも成分母音の特徴を失うことなく得られている。



/a/(M1) + /i/(M2)

(a) 自己共分散関数 $R(\tau;k)$. (b) $\tau = 0, T_1, T_2$ における $\sqrt{R(\tau;k)}$.

Fig. 3-2 /a/ (M1) + /i/ (M2) に対する自己共分散関数と分離スペクトルパターン(破線は単独母音のスペクトルパターン).



/a/(M1) + /u/(M2)

(a) 自己共分散関数 R(τ;k). (b) τ = 0, T1, T2 における √R(τ;k).

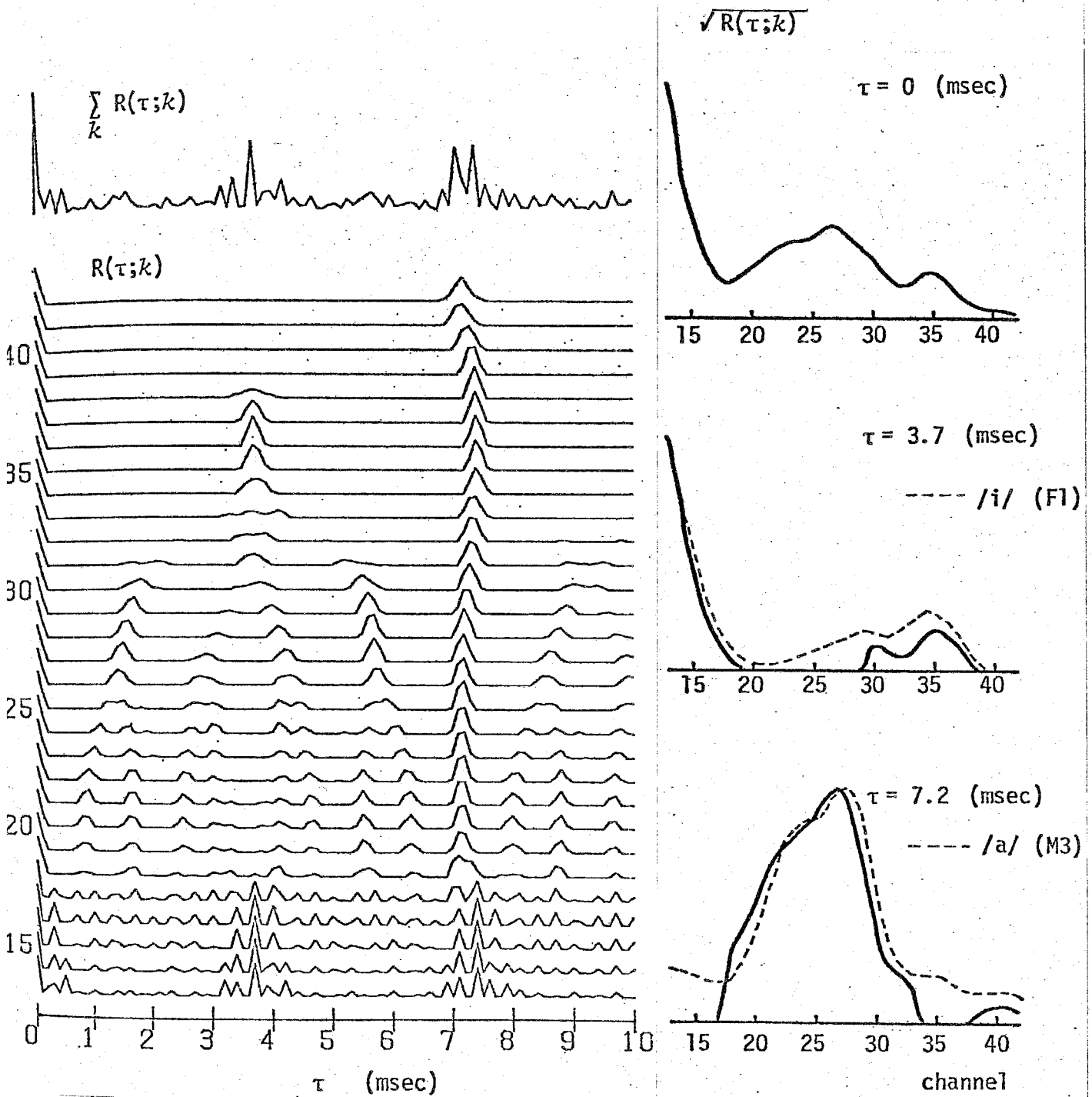
Fig. 3-3 /a/ (M1) + /u/ (M2) に対する自己共分散関数と分離スペクトルパターン (破線は単独母音のスペクトルパターン).

（2）．各成分母音のピッチが近い場合（一方のピッチが他方のピッチの整数倍に近い場合も含む）．

Fig．3－4に，一方のピッチが他の成分母音のピッチの約2倍となっている例，男性（M3）の/a/と女性（F1）の/i/の混合音声に対する結果を示す。女性母音のピッチは，13～17チャネルおよび34～38チャネルにかけてのピークより $T_1 = 3.7$ msec と検出される。$t = 7.2$ msec 付近では，$T_2 \fallingdotseq 2T_1$ の関係から全チャネルにわたりピークが出現しているが，対称性に注目すれば，19～28チャネルのピークから $T_2 = 7.2$ msec と男性母音のピッチが検出できる。この結果，Fig．3－4（b）では，$t = T_1, T_2$ の各位置から成分母音のスペクトルパターンが分離されていることがわかる。

（3）．同一母音でピッチも近い場合。

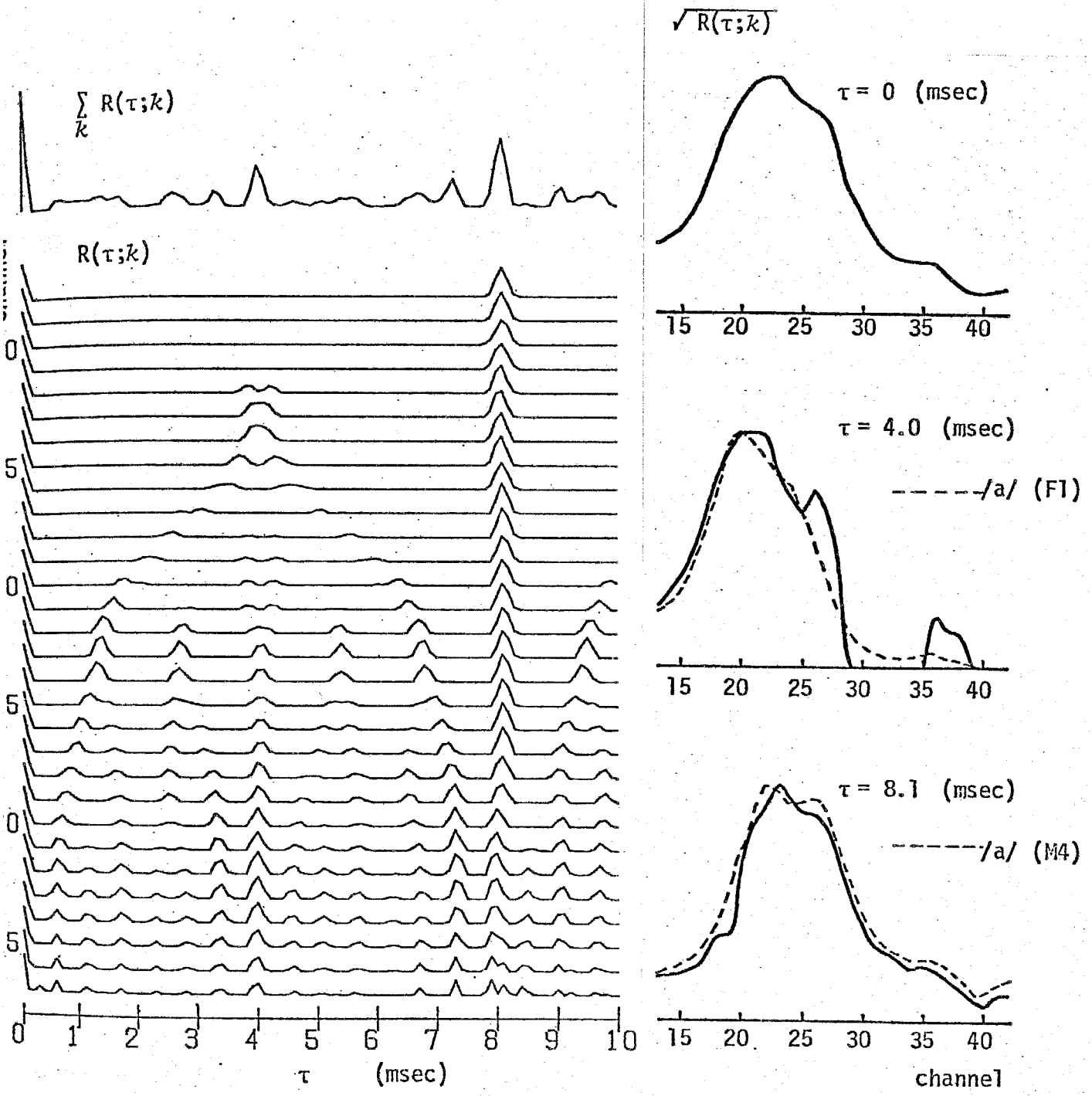
前項では，ピッチは互いに近かったものの成分母音が異なるため，その周波数成分が異なり，各特徴の出現するチャネルが異なるので，ピッチ検出がうまくできたと考えられる。そこで，互いにピッチも近く，且つ同一母音からなる混合音声の分離を試みた。その一例として，男性（M4）の/a/と女性（F1）の/a/の混合音声に対する結果を，Fig．3－5に示す。Fig．3－5（a）から，女性母音のピッチは $T_1 = 4.0$ msec と簡単に検出でき，また男性母音のピッチも，パターンの対称性に注目すれば，21～32チャネルから $T_2 = 8.1$ msec と検出できる。Fig．3－5（b）では，同一母音ながら各話者の個人的特徴（同図中の破線のパターン）をも失うことなく，それぞれのスペクトルパターンが分離されていることがわかる。



/a/(M3) + /i/(F1)

(a) 自己共分散関数 $R(\tau;k)$. (b) $\tau = 0, T_1, T_2$ における $\sqrt{R(\tau;k)}$.

Fig. 3-4 /a/(M3)+/i/(F1)に対する自己共分散関数と分離スペクトルパターン(破線は単独母音のスペクトルパターン).



/a/(M4) + /a/(F1)

(a) 自己共分散関数 $R(\tau;k)$. (b) $\tau = 0, T_1, T_2$ における $\sqrt{R(\tau;k)}$.

Fig. 3-5 /a/ (M4) + /a/ (F1) に対する自己共分散関数と分離スペクトルパターン (破線は単独母音のスペクトルパターン).

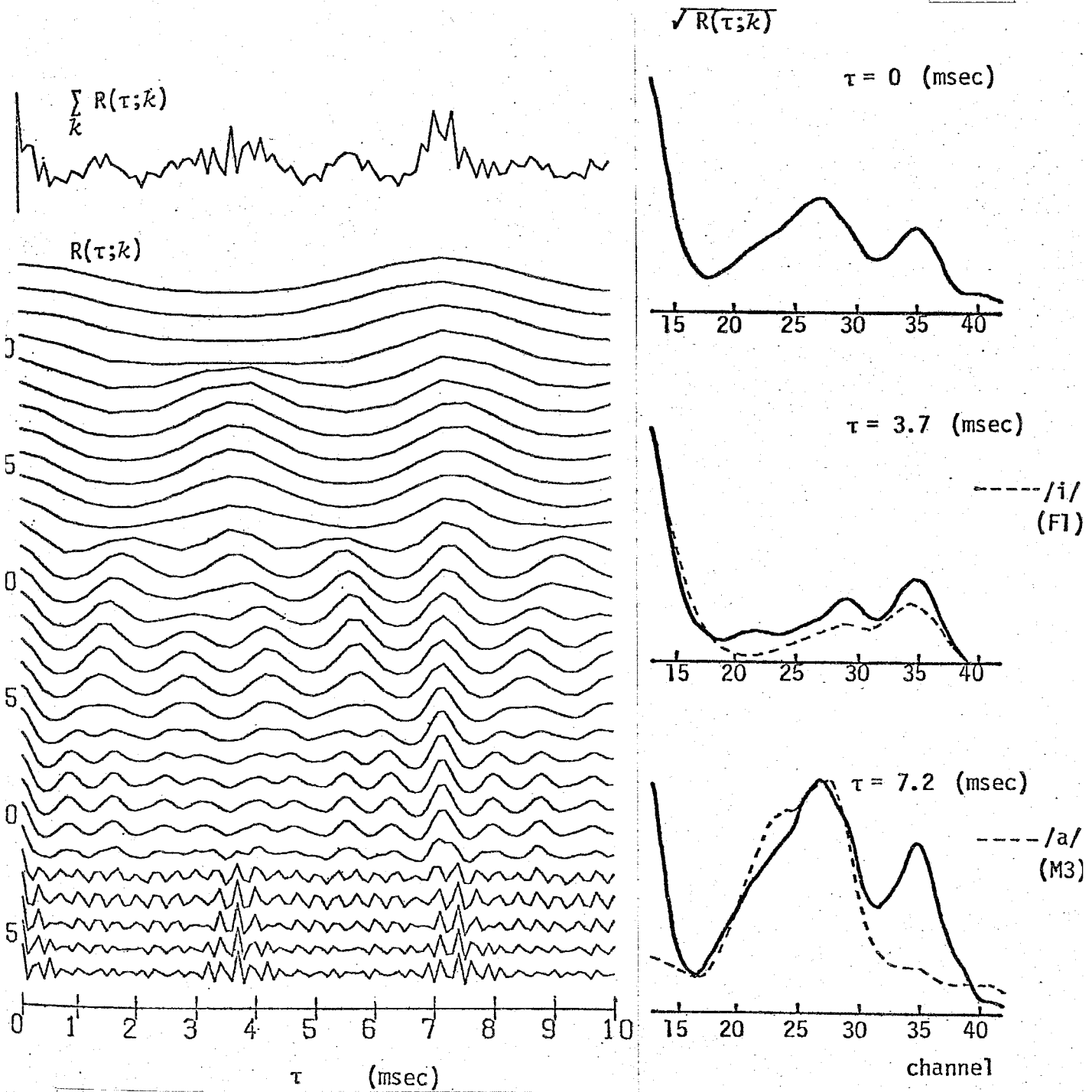
3.4 検討

本章で提案した聴き分け機構のモデルは，聴覚神経では音声情報は，振動に同期したパルスの形で伝えられることを考慮し，基底膜出力をパルス列に変換してから自己共分散関数を求めるというものであるが，このパルス列への変換操作が聴き分け能力にどのように影響を与えているかを検討してみる。

自己共分散関数の性質からすれば，基底膜出力をパルス列に変換せずとも成分母音のピッチ検出は可能である。そこで Fig. 3－4に示したと同一の入力（/a/（M3）+/i/（F1））について，半波整流波形そのものの自己共分散関数を求め，同一の手続きで成分母音のスペクトルパターンの分離を試みた。その結果を，Fig. 3－6に示す。Fig. 3－6（a）に見られるように，各成分母音のピッチに対応したピークのすそが広がり，ピーク位置検出の精度は低下するが，ほぼ $T_1 = 3.7$ msec，$T_2 = 7.2$ msec と同定できる。Fig. 3－6（b）では，女性母音 /i/ のスペクトルパターンは，Fig. 3－4（b）よりもむしろ忠実に再現されているが，$\sqrt{R(T_2; k)}$（$T_2 = 7.2$ msec）の図では，/i/ のスペクトルパターンが重畳し，35チャネルと13チャネルに2つの大きなピークが出現してしまい，母音 /a/ のスペクトルパターンの分離には成功していない。

以上の結果，パルス列への変換が，分離スペクトルパターンに悪影響を及ぼすことなく，共分散関数のピークの幅を狭め成分母音のピッチ検出を容易にし，且つ，スペクトルパターンの分離能力を向上させることが分った。

実際の神経系では，振幅情報はパルス高ではなく パルスの発生確率（パルス頻度）という形で伝送される。しかしそのような場合でも，相関をとるデータの長さが十分にあれば，同様にスペクトルパターンは分離できるであろう。



/a/(M3) + /i/(F1) (without pulse generation)

(a) 自己共分散関数 $R(\tau;k)$ (b) $\tau = 0, T_1, T_2$ における $\sqrt{R(\tau;k)}$.

Fig. 3-6 半波整流操作のみの自己共分散関数と分離スペクトルパターン.
入力は /a/ (M3) + /i/ (F1).

3.5 むすび

人間の聴き分け機構を探るために，聴き分けのキーポイントはピッチの違いにあると仮定し，神経系におけるパルス相関モデルを構成し，混合音声の聴き分けをシミュレートした。2名の話者により同時発声された持続母音からなる混合音声について，成分母音のピッチ抽出ならびにスペクトルパターンの分離抽出を試みた結果，このモデルでは原理的に不可能な成分音のピッチが全く同一か，他方のピッチの整数倍に完全に重なる場合を除けば，成分母音のスペクトルパターンを個人性情報をもあまり損なうことなく分離することができた。

本研究では，母音についてのみ聴き分けを試みたが，子音に対しては，ピッチを持つ有声子音ならばこの方法でスペクトルパターンは分離できるはずである。しかし，無声子音に対してはピッチ成分が無いので不可能である。また，有声子音に対しても，時間変動が大きい場合には，共分散関数を求めるために使用するデータ長（時間）等，考慮せねばならない。さらに，今回の実験ではピッチの検出は視察により行なったが，実際に音声認識装置に応用するためには自動的にこれを検出する方法についても検討する必要がある。

人間が実際に聴き分けを行なっている時には，両耳聴の効果もあり，また Syntax や Semantics などの高次中枢系の作用の助けを借り，特徴分離の不完全さを補なっている。人間の聴き分け機構の解明には，このような下位の神経系での特徴分離実験のみならず，音声理解系の研究も合わせて行なう必要がある。

また，この実験を通じて，パルスでの情報伝送が聴き分け能力の向上に役立つことが分ったが，実際の神経系でもパルスにより情報が伝送されていることから，それが人間の認識能力に反映されている場合もあると思われる。

第4章　単母音認識

4.1　まえがき

日本語は母音中心型の言語であり，日本語の連続音声認識システムを構成しようとする場合，その音韻認識部における母音認識のしめる役割は特に大きいものとなろう。そこで，不特定話者の母音を認識するすぐれたシステムを開発する必要がある。本章では，基底膜モデルから得られるスペクトルパターンを用いた不特定話者の単母音認識システムについて述べる。

不特定話者を対象とする認識システムを構成しようとする場合，話者としては成人男女及び子供を考えれば十分であろう。そこで，成人男性32名，同女性25名，10～12才の男子44名，同女子16名，合計117名の発声した単母音のスペクトルパターンを詳しく観察して，認識システムを構成した。

本章で提案する母音認識システムは，第1段階として，入力スペクトルパターンと参照パターンとの距離を算出しその距離をもとに一次候補を選択し，次に第2段階で入力スペクトルパターンの形態的特徴が一次候補に上げられたカテゴリの特徴と合うかを検証し，最終判断を下すという2段階の構成となっている（Fig. 4-1 参照）。話者の属性（年令・性別）による違いを考慮するために「属性クラス」として「男性」「女性」「子供」の3クラスを，また5母音に対応する「音韻クラス」として /a/, /i/, /u/, /e/, /o/ の5クラスを考え，属性クラス別音韻クラスの合計15カテゴリへの分類を実行する。以下，この15カテゴリを，「男性」/a/ なら [MA]（Male /a/），「女性」/i/ なら [FI]（Female /i/），「子供」/u/ なら [CU]（Child /u/）等と表現する。また，本システムは，入力音声からピッチ抽出等を行ない話者の属性（年令・性別）を決定してから音韻の識別を実行するのではなく，話者の属性と音韻性をまとめて取り扱うという特徴を持っている。

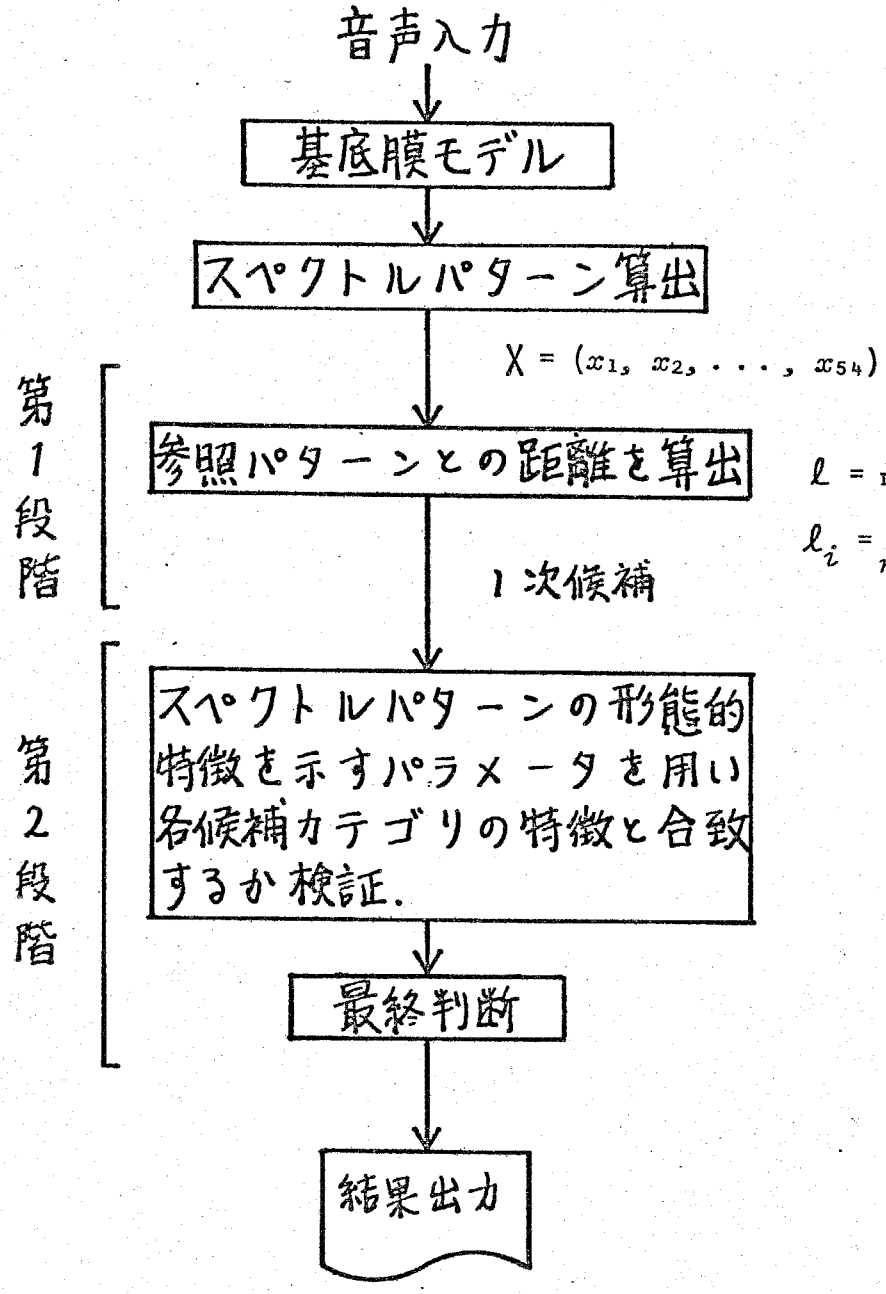


Fig. 4-1 母音認識処理の概略.

4.2 スペクトルパターン

基底膜モデル全544チャネルの出力の短時間平均2乗振幅を用いて，スペクトルパターンを次のように定義する。スペクトルパターンを，各チャネル出力に対応する54次元のベクトル $\mathbf{X}=(x_1, x_2, \cdots x_{54})$ と表現する。各元 x_k は，k チャネルの基底膜出力を $f_k(n)$ として（4－1）式で定義する。

$$x_k = \left(\frac{\sum_{n=0}^{N-1} f_k^2(n)}{\sum_{l=1}^{54} \sum_{n=0}^{N-1} f_l^2(n)} \right)^{1/2} \qquad (4-1)$$

$N=512$　　平均時間 $NT=12.8$ msec

（T：サンプリング間隔 25 μsec）

本システムでは，フレーム長12.8 msec，フレーム周期12.8 msecとしてスペクトルパターンを求め，認識に利用する。

Fig.4－2に，成人男性，同女性，子供の発声した5母音のスペクトルパターンの様子を示す。Fig.4－2は，各10名ずつのスペクトルパターンを最大値を一致させて重ね書きしたものであるが，各母音でそれぞれ特徴的な形を示していることがわかる。また，個人差による変動もかなり小さいことがわかる。

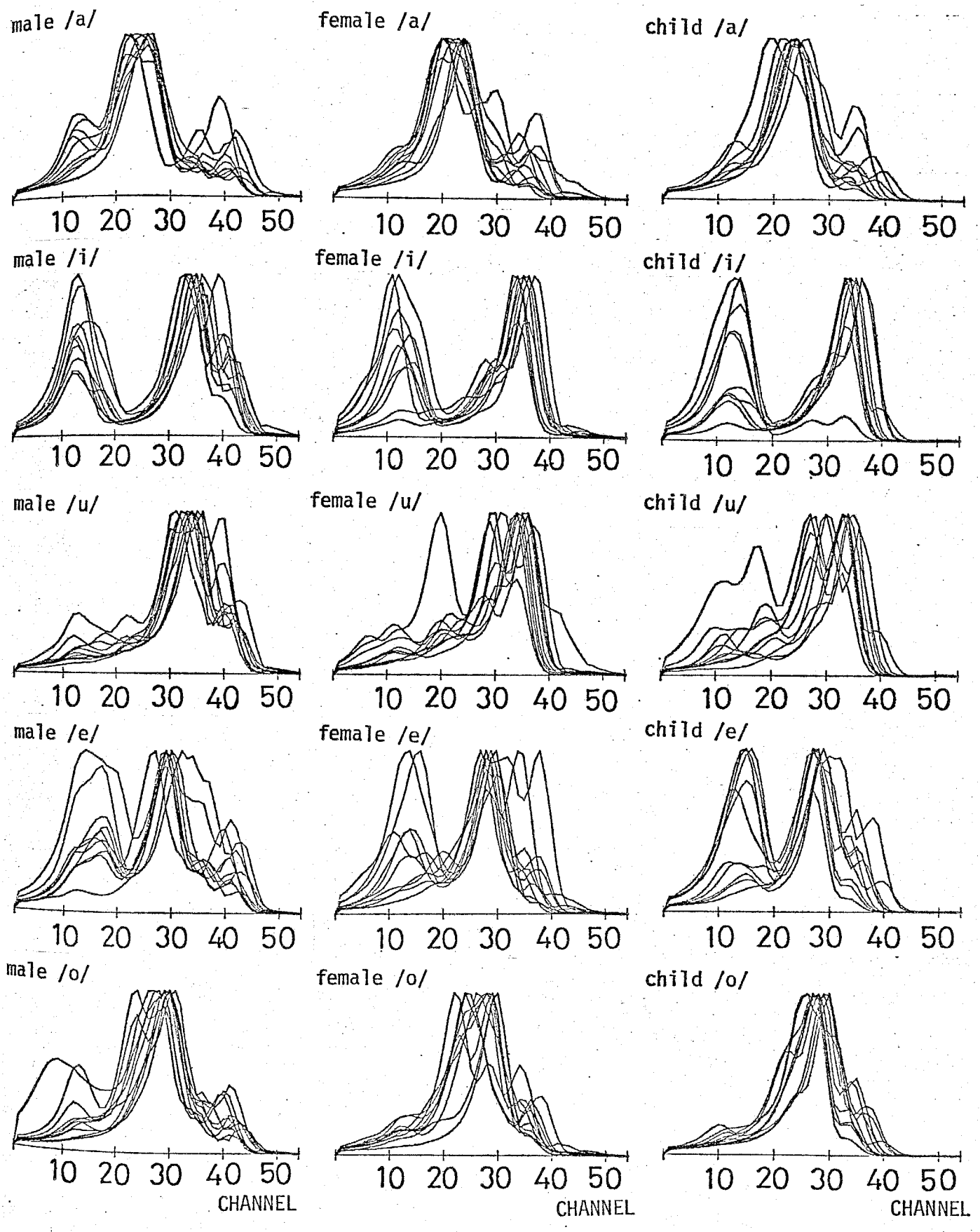


Fig. 4-2 成人男性，同女性，子供 各10名の発声による
日本語5母音のスペクトルパターン.

4.3 参照パターン

第1段階の処理で用いる参照パターンの作成法について述べる。まず，全サンプル（成人男性32名，同女性25名，10～12才の男子44名，同女子16名の発声した単母音，合計585個）の定常部分の1フレームのスペクトルパターンを求め，これらを視察により求めたスペクトルパターン上のピッチのピークの出現チャネルにより，次の3つの属性クラスに分類する。この分類は，ピッチの違いにのみよるもので，各属性クラスのラベルと実際の話者の年令・性別とは必ずしも一致しない。

ピッチのピーク位置

40ch以後（周波数150Hz以下）	「男性」
36～40ch（同250～150Hz）	「女性」
36ch以前（同250Hz以上）	「子供」

次に，5つの音韻クラスをこの属性クラス別にした合計15カテゴリの各々に対して，典型的だと思われる10個のスペクトルパターン（サンプルパターン $X^1 \sim X^{10}$）を視察により選出し，それらをもとに以下のアルゴリズム①～④により参照パターンを作成した。

①．各々のカテゴリに対して，10個のサンプルパターンの中から任意に1つ（X^1）を選び，これを仮参照パターン $Y^1 = (y_1, y_2, \cdots y_{54})$ とし，$M=1$ から $M=9$ まで ②，③の手順を繰り返して，仮参照パターン Y^{10} を求める。

②．別の任意のサンプルパターン $X^{M+1} = (x_1^{M+1}, x_2^{M+1}, \cdots x_{54}^{M+1})$ と仮参照パターン $Y^M = (y_1^M, y_2^M, \cdots, y_{54}^M)$ との間で，次の3種の距離 d_i（$i=0, \pm 1$）を（4－2）式に従い算出する。

$$d_i = \sum_{n=2}^{42} \left(y_n^M - x_{n+i}^{M+1} \right)^2 \qquad i = 0, \pm 1 \qquad (4-2)$$

(4-2)式は，サンプルパターン X^{M+1} を±1チャネルのシフトを施したものを含めてパターン Y^M とマッチングをとることを表わしているが，これは各カテゴリ内でのピッチや個人差による変動を少しでも小さくするために実施するものである。また，一般の会話音声信号は 100 Hz 以下の成分は持たないことから，この距離 d_i の算出には 43 チャネル（CF 111 Hz）までの値を用いる。

③. ②で求めた d_i（$i = 0, \pm 1$）の中で最小値を示す i-ch シフトを施したサンプルパターンで 仮参照パターン Y^M を（4-3）式に従い修正し，Y^{M+1} を得る。

$$y_n^{M+1} = \frac{1}{M+1} \left(M y_n^M + x_{n+i}^{M+1} \right) \qquad (4-3)$$

$n = 2 \sim 42$，i：d_i が最小となる i

④. ①～③の手順により各カテゴリに対して1つずつ求められた合計15個の仮参照パターン Y^{10} の各々と，585個の全スペクトルパターン X との間の距離 l を，（4-4）式により求めた結果，スペクトルパターンが属すべき音韻クラス（例えば [MA] [FA] [CA] が同一音韻クラス）に対する l が最小とならなかったスペクトルパターンを用い，上記②，③の手順により，そのスペクトルパターンが属すべき音韻クラスの中で最小の l を持つカテゴリの仮参照パターンを修正し，最終的な参照パターンとする。

$$\left.\begin{aligned} l &= \min. \{ l_i \mid i = 0, \pm 1 \} \\ l_i &= \sum_{n=2}^{42} \left(\log y_n^{10} - \log x_{n+i} \right)^2 \end{aligned}\right\} \qquad (4-4)$$

(例). 成人男性 /a/ のスペクトルパターン※と 各カテゴリの仮参照パターン Y^{10} との間の距離 ℓ を算出したところ，次のような結果となったとする。

ℓ[MO] = 0.69
○ ℓ[FA] = 0.70
○ ℓ[MA] = 0.73
ℓ[ME] = 0.98
ℓ[FO] = 1.01
ℓ[CO] = 2.18
ℓ[FE] = 3.34
○ ℓ[CA] = 5.66
ℓ[CU] = 8.67
ℓ[FU] = 9.01
ℓ[CE] = 10.85
ℓ[CI] = 11.32
ℓ[FI] = 13.93
ℓ[MU] = 14.62
ℓ[MI] = 15.83

ℓ[MO] が最小となり，スペクトルパターンが属すべき音韻クラスの ℓ（ℓ[MA]，ℓ[FA]，ℓ[CA] のいずれか）が最小となっていない。この時，音韻クラス /a/ の中で最小の ℓ を示すものは，ℓ[FA] であるので，[FA] の仮参照パターン Y^{M}[FA]（$M \geq 10$，すでに他のスペクトルパターンで修正されている場合は $M > 10$）を※を使い②，③の手順で修正する。

以上の手順により求められた参照パターンを Fig. 4-3 に示す。

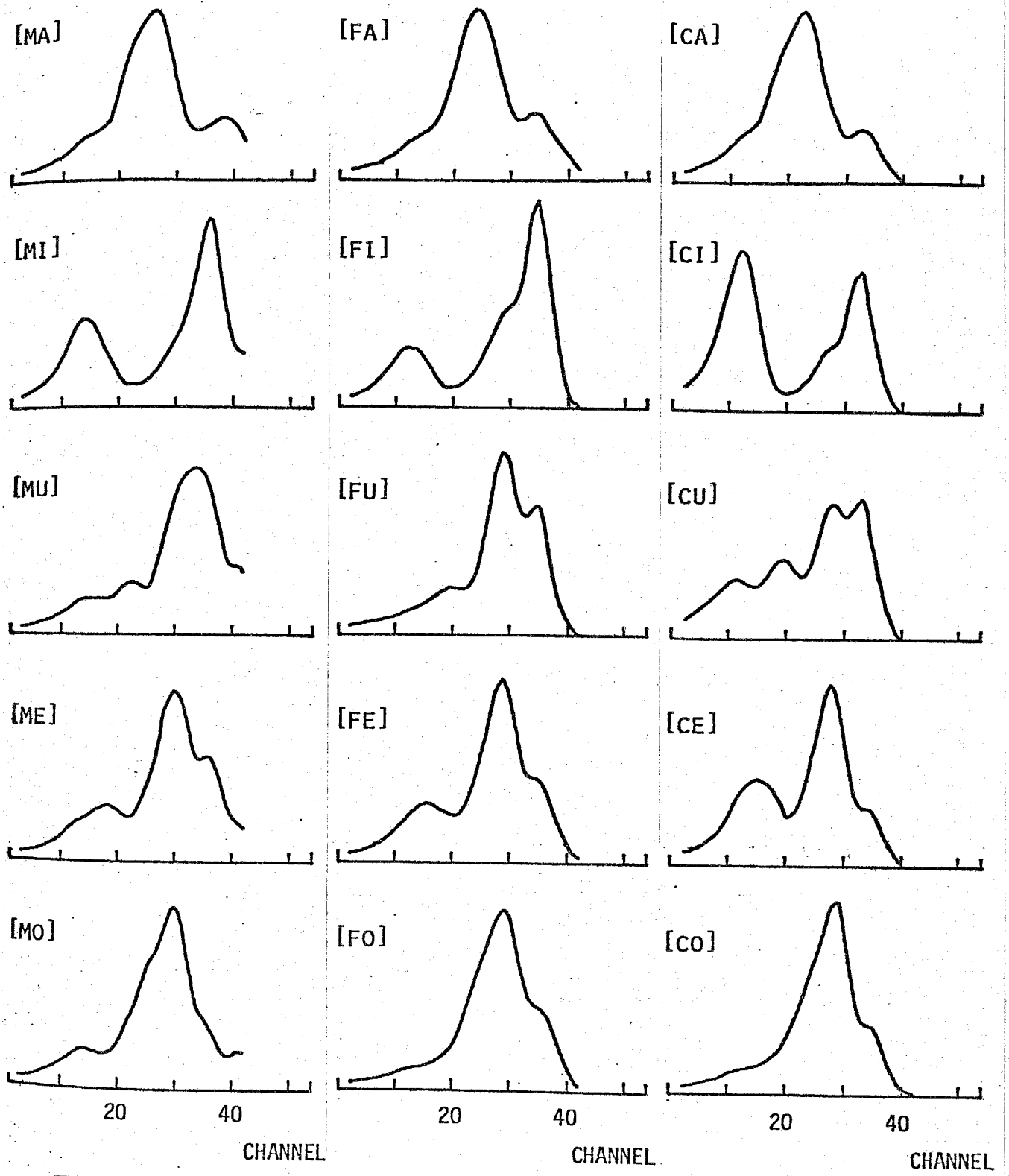


Fig. 4-3 参照パターン.

4.4 特徴パラメータ

Fig. 4-2に現われているように各母音のスペクトルパターンは，ピークの位置やディップの位置等，それぞれ特徴的な形を示す。この節では，認識処理の第2段階で用いるスペクトルパターンの形態的特徴を記述するパラメータについて説明する。

(1). ピークの位置とピークの大きさの順序関係.

スペクトルパターン上に現われるすべてのピークを，ピークの大きい順に番号づけ，その出現チャネルを P_1，P_2，$\cdots$ P_n と表現する。母音の認識にはフォルマントの様にエネルギーの集中した部分が重要なので，ピーク検出の際には，基底膜のQ値の低さを補うために Fig. 4-4 に示す側抑制フィルタをスペクトルパターンにかける。側抑制をかけた後のスペクトルパターンを $\hat{X} = (\hat{x}_1, \hat{x}_2, \cdots \hat{x}_{54})$ とすると，$\hat{x}_k$ は（4-5）式で求められる。

$$\hat{x}_k = 2x_k - 0.3(x_{k-1} + x_{k+1}) - 0.2(x_{k-2} + x_{k+2}) \quad (4-5)$$

ピークの判別は，$(\hat{x}_p - \hat{x}_{p-1}) > 0$，かつ $(\hat{x}_{p+1} - \hat{x}_p) < 0$ なるチャネル p をもってピーク位置とし，側抑制をかける以前の値 x_p をもってピークの大小を決定する。スペクトルパターン上に現われるピークを形成する原因は，ピッチ周波数やフォルマント周波数にあるが，基底膜モデルはQ値が低いので，たとえ側抑制フィルタをかけたとしても，2つのフォルマント成分が存在するにもかかわらず1つのピークしか検出できない場合もある。また，ピッチ周波数によるピークかフォルマントによるピークか区別できない場合も生じる。故に，スペクトルパターン上に出現するピークは，その生成原因にかかわらずすべてを検出する。

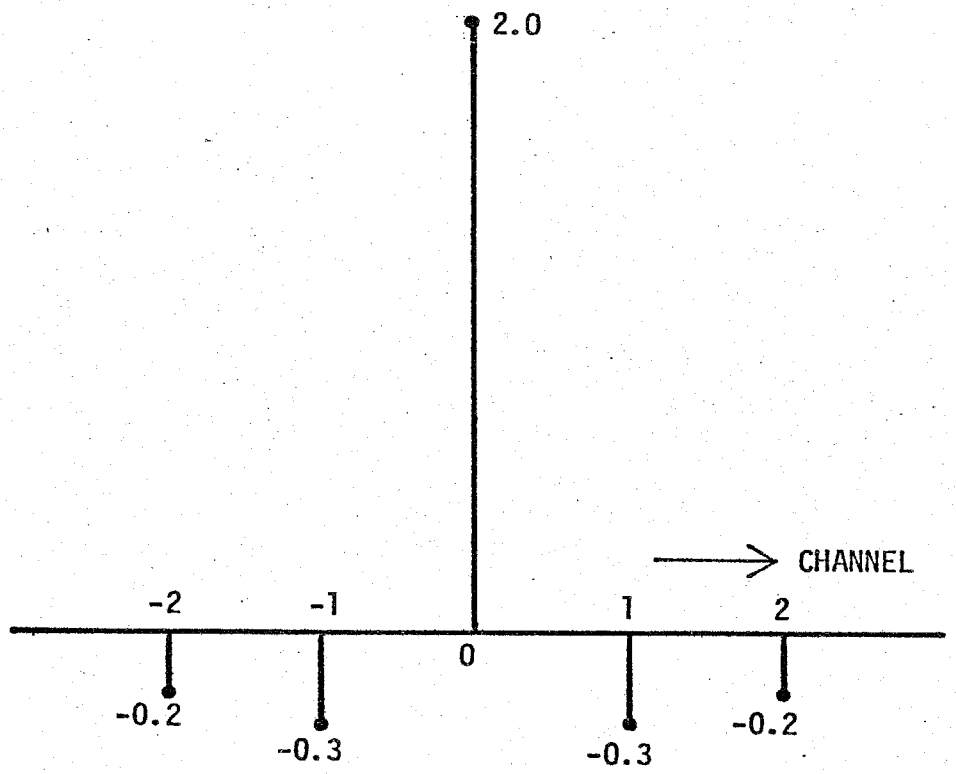


Fig. 4-4 側抑制フィルタ.

(2). ディップの位置.

スペクトルパターン上のディップを その出現チャネルの小さい順に番号づけ, その出現チャネルを D_1, D_2, ・・・ D_n と表現する。ピークは母音の音韻性を決定するのに重要であるのに対し, ディップはスペクトルパターンの概略形を把握するために用いるので, 側抑制フィルタは使用せず, $(x_\alpha - x_{\alpha-1}) < 0$ かつ $(x_{\alpha+1} - x_\alpha) > 0$ なるチャネル α をもってディップの位置とする。

(3). 最大ピークの幅.

最大ピークの幅を表現するパラメータとして，次の3つのパラメータ W-，W+，W を用いる。最大ピーク（出現チャネル P_1 ）の値 x_{P_1} の1／2を閾値として，値 x が x_{P_1}／2 以下になるまでのチャネル数をもって，W-，W+，W を Fig. 4-5 のように定義する。W- は最大ピークの高周波数側へのひろがりを，W+ は低周波数側へのひろがりを，W (= W- + W+) は，ピークの幅をそれぞれ表現する。このパラメータは，最大ピークの出現チャネルが互いに近く，2つのフォルマント F_1，F_2 が接近していてスペクトルパターン上では分離されにくい /a/ と /o/ の区別に役立つ（Fig. 4-11参照）。

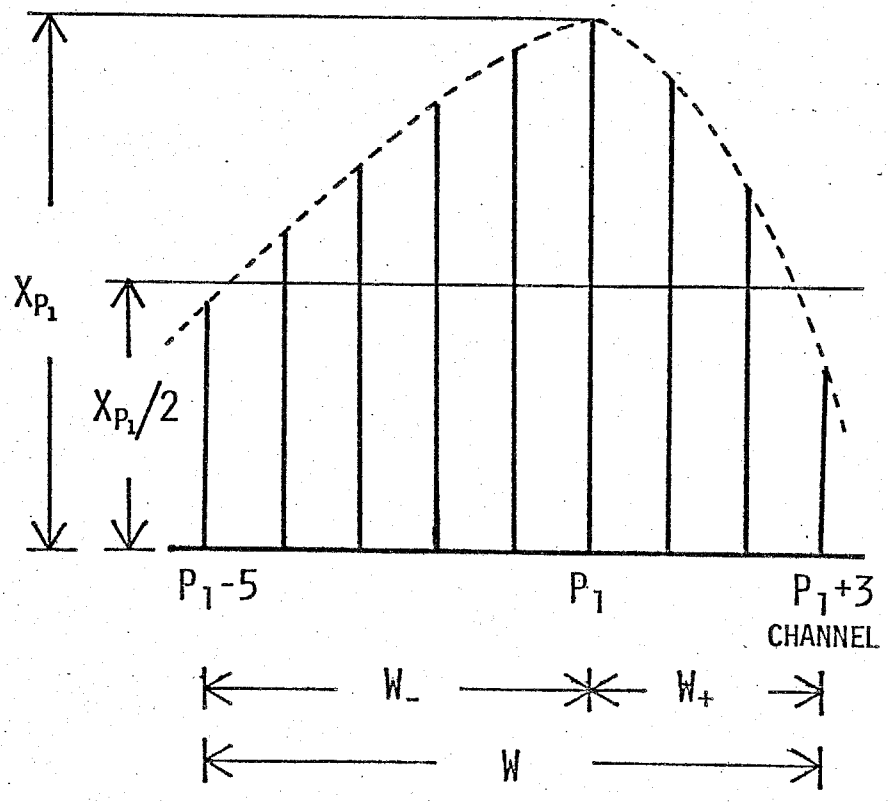


Fig. 4-5 W-，W+，W の定義.

(4). ピーク間の距離.

ある区間をおいて存在する2つのピーク間の距離を示すパラメータとして(4-6)式で定義される量を導入する。

$$dis[m, n] = \min\{P_\ell \mid P_\ell \geqq n\} - \max\{P_k \mid P_k \leqq m\} \quad (4-6)$$

つまり, $dis[m, n]$は, n-ch以後のピークの中で最もn-chに近いピークと, m-ch以前のピークで最もm-chに近いピーク間の距離をチャネル数で表わしたものである。このパラメータは, スペクトルパターンが2つの大きなピークを持つ/i/と/e/の区別に役立つ(Fig. 4-15参照)。

Fig. 4-6 ~ Fig. 4-10に各母音のピークとディップの分布ヒストグラムを, Fig. 4-11に/a/と/o/におけるP_1と$(W_- - W_+)$の関係を, Fig. 4-12 ~ Fig. 4-14に/i/, /u/, /e/におけるP_1の分布ヒストグラムを, Fig. 4-15に/i/と/e/における$di[m, n]$の様子を示す。Fig. 4-15においては, [MI]については$dis[18, 28]$, [FI]は$dis[15, 28]$, [CI]は$dis[16, 26]$の分布を示し, また[ME]は$dis[20, 27]$, [FE]は$dis[20, 26]$, [CE]は$dis[17, 24]$の分布を示す。これら/i/及び/e/における$dis[m, n]$は, ほとんどすべてが隣接するピーク間の距離である。

以上の4種のパラメータを用いて 認識処理の第2段階での各カテゴリの検証アルゴリズムを構成する。

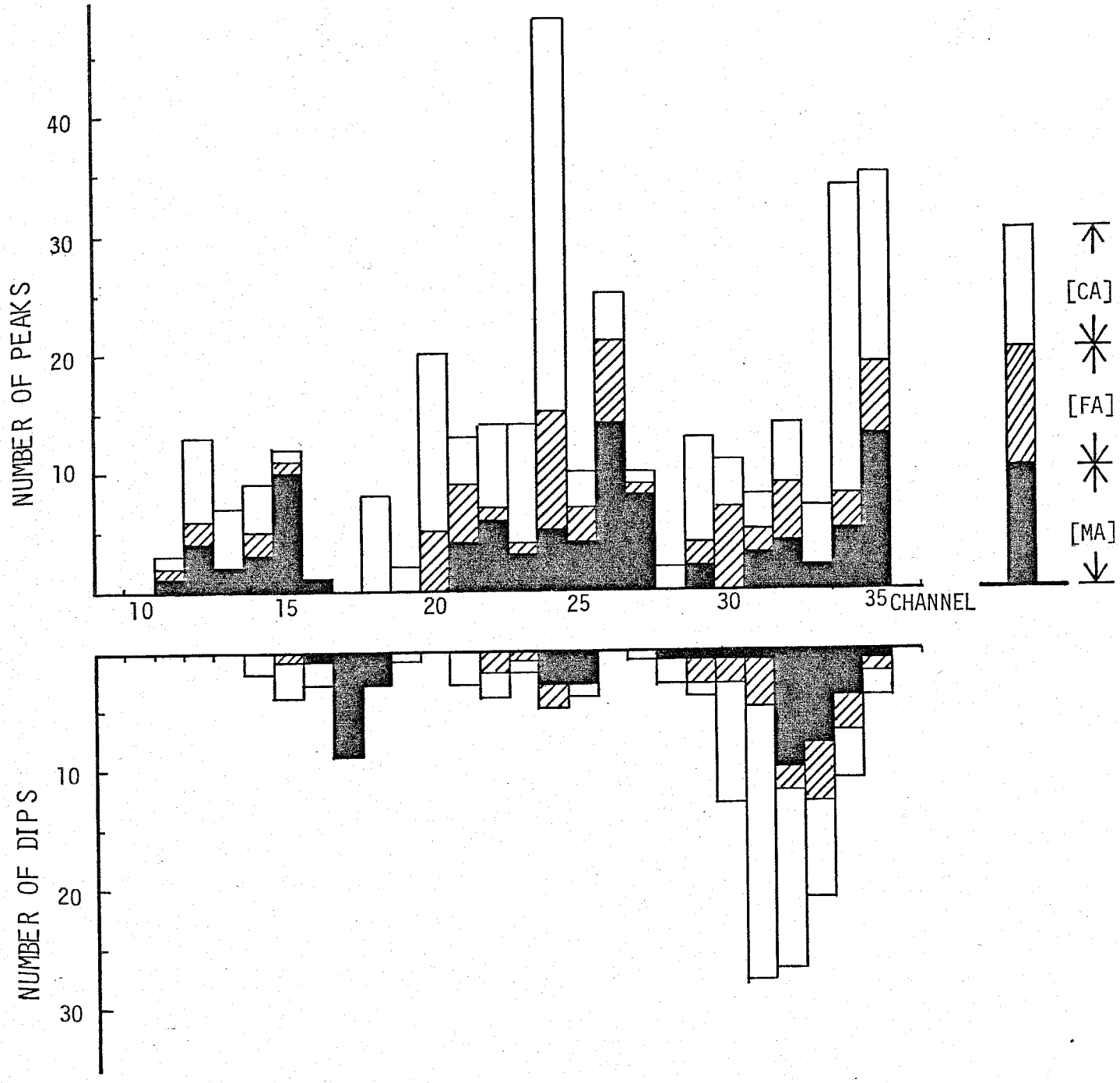


Fig. 4-6 /a/ におけるピークとディップの分布ヒストグラム.

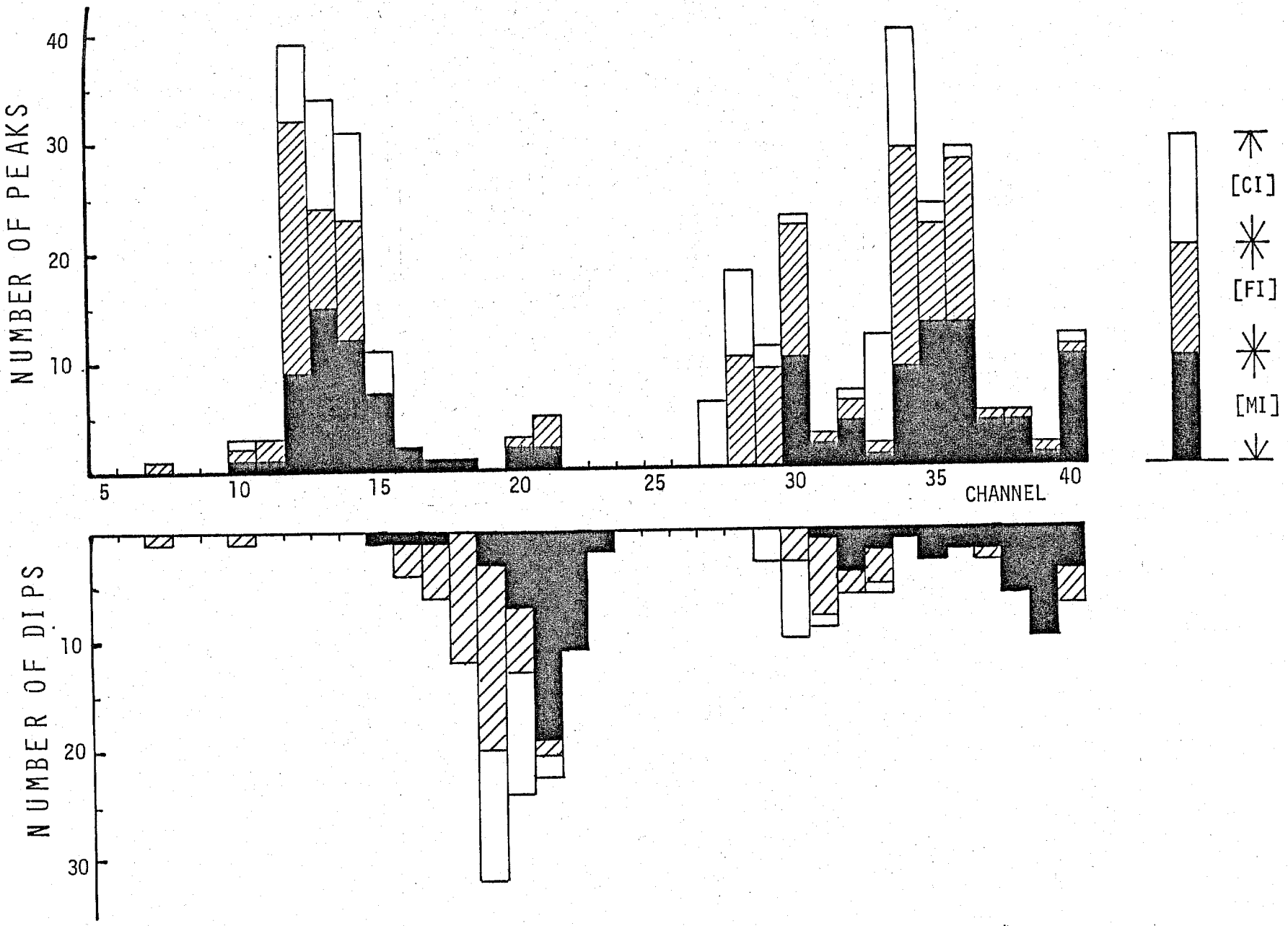


Fig. 4-7　ノイノにおけるピークとディップの分布ヒストグラム.

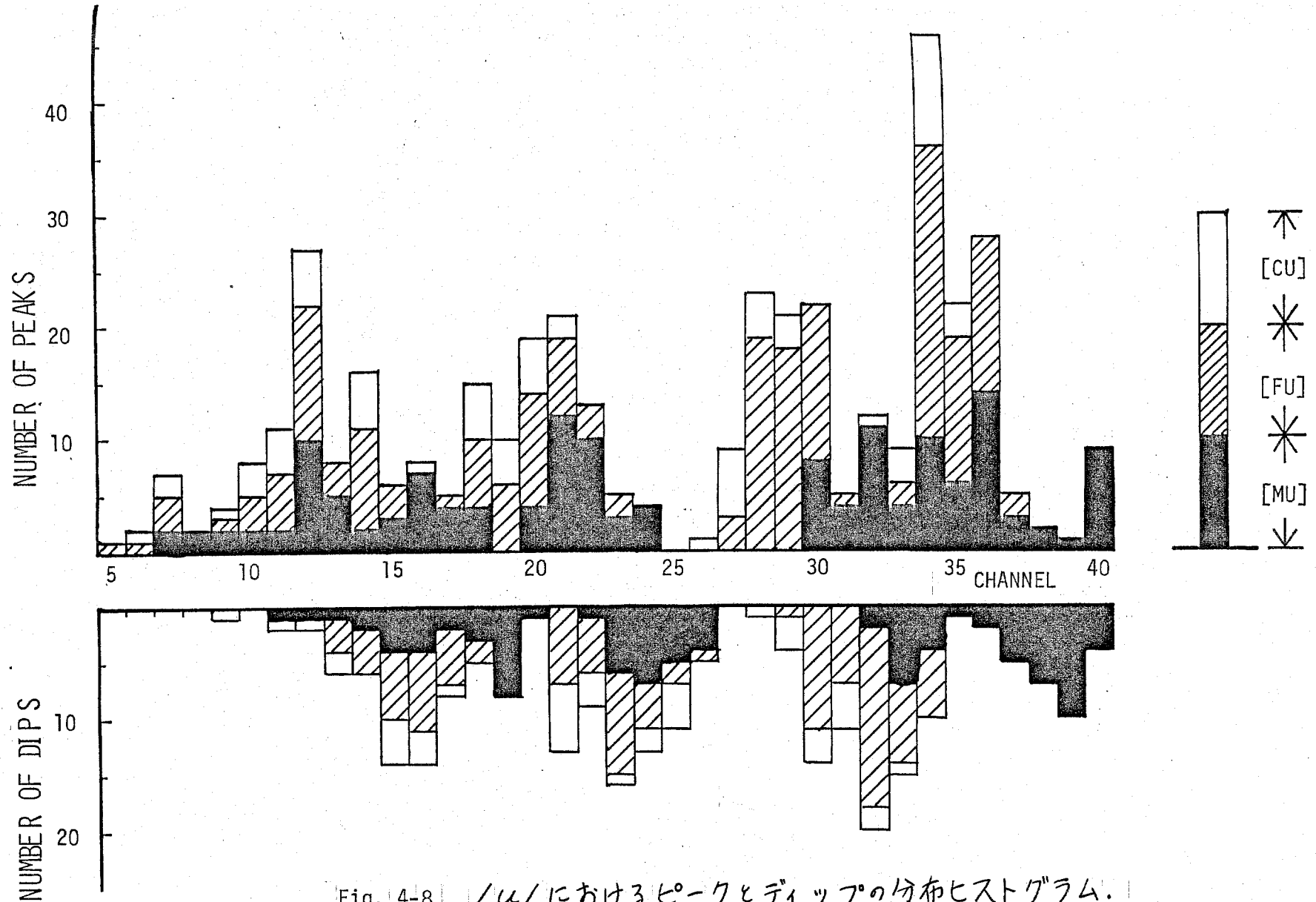


Fig. 4-8 /u/におけるピークとディップの分布ヒストグラム.

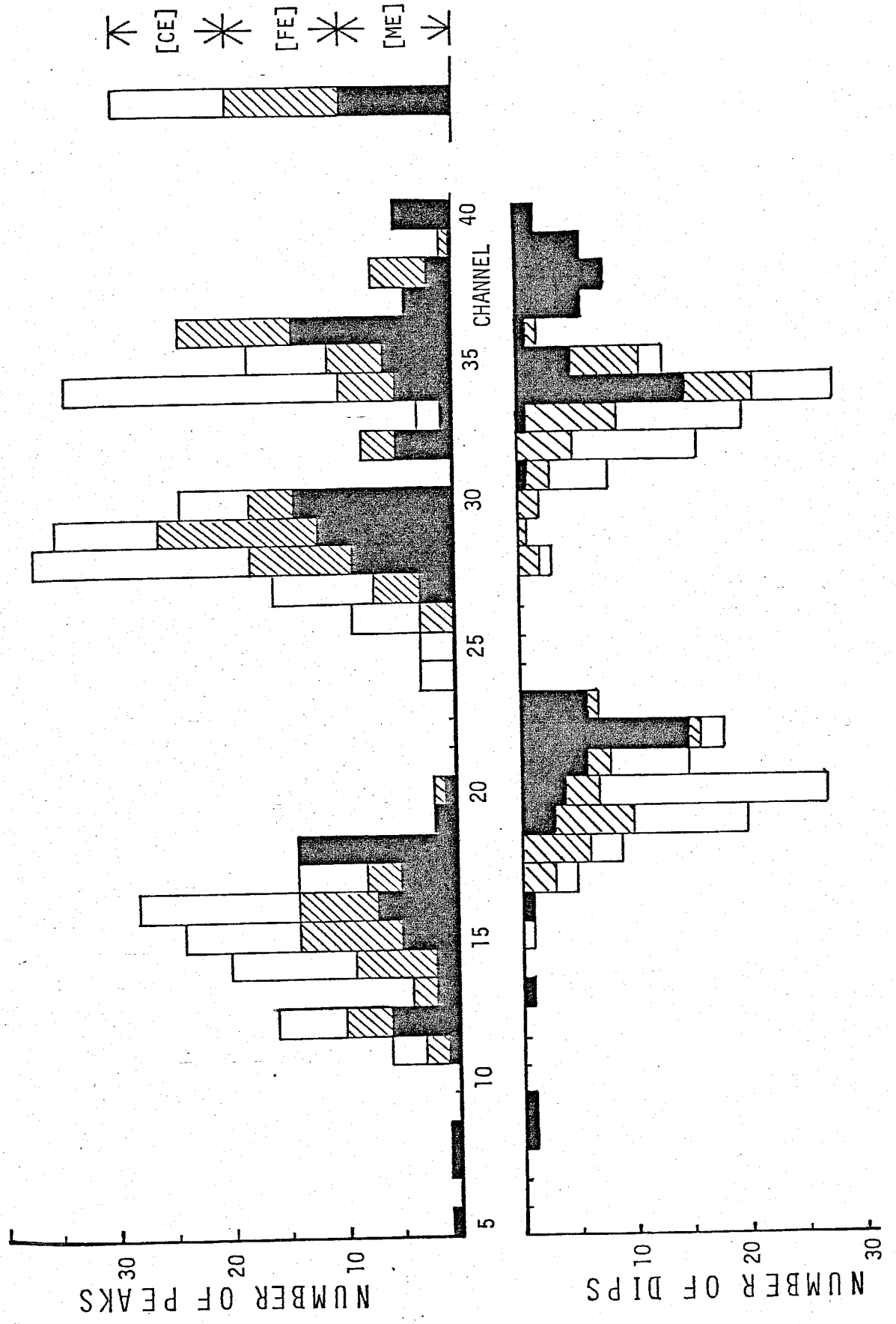


Fig. 4-9 /e/ におけるピークとディップの分布ヒストグラム.

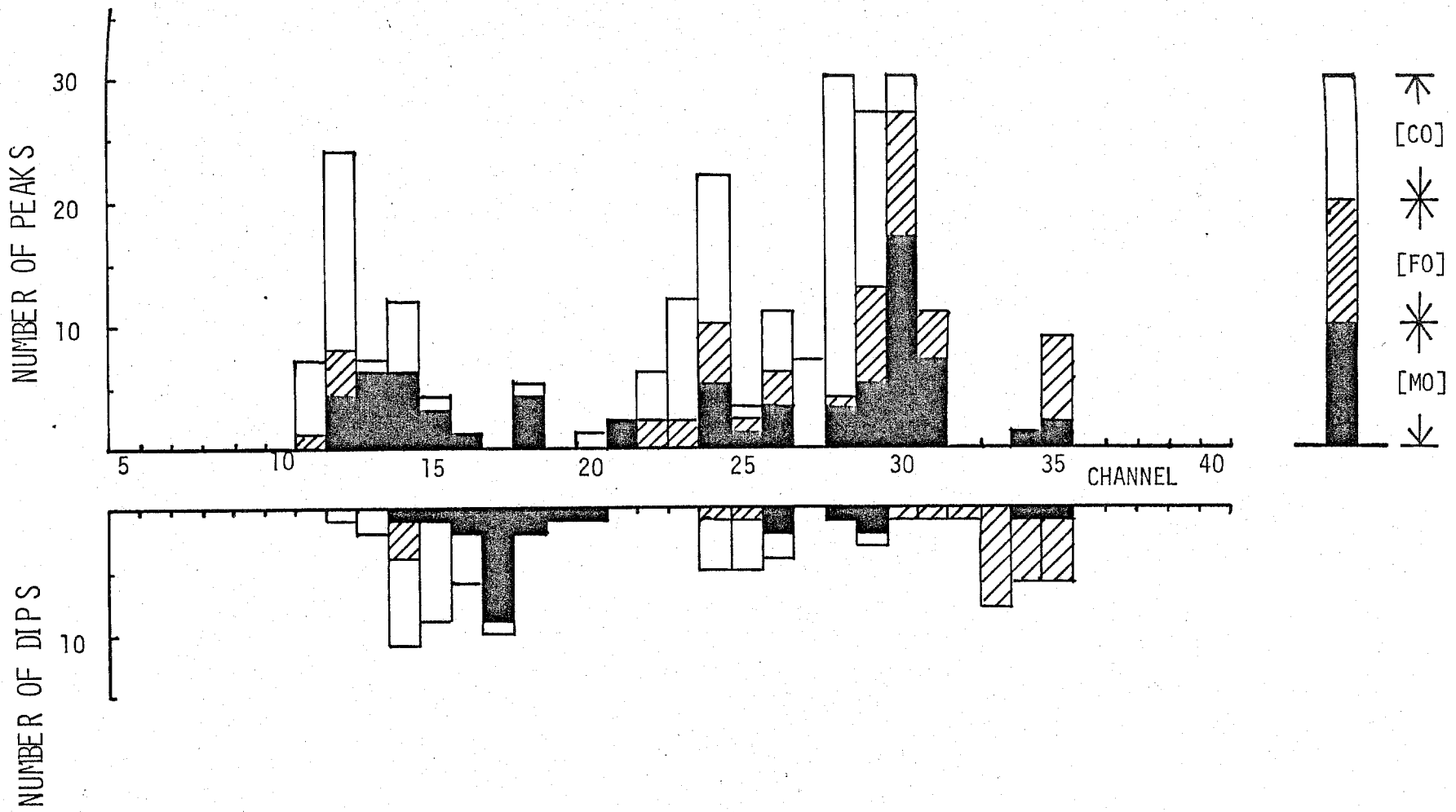


Fig. 4-10 /O/におけるピークとディップの分布ヒストグラム.

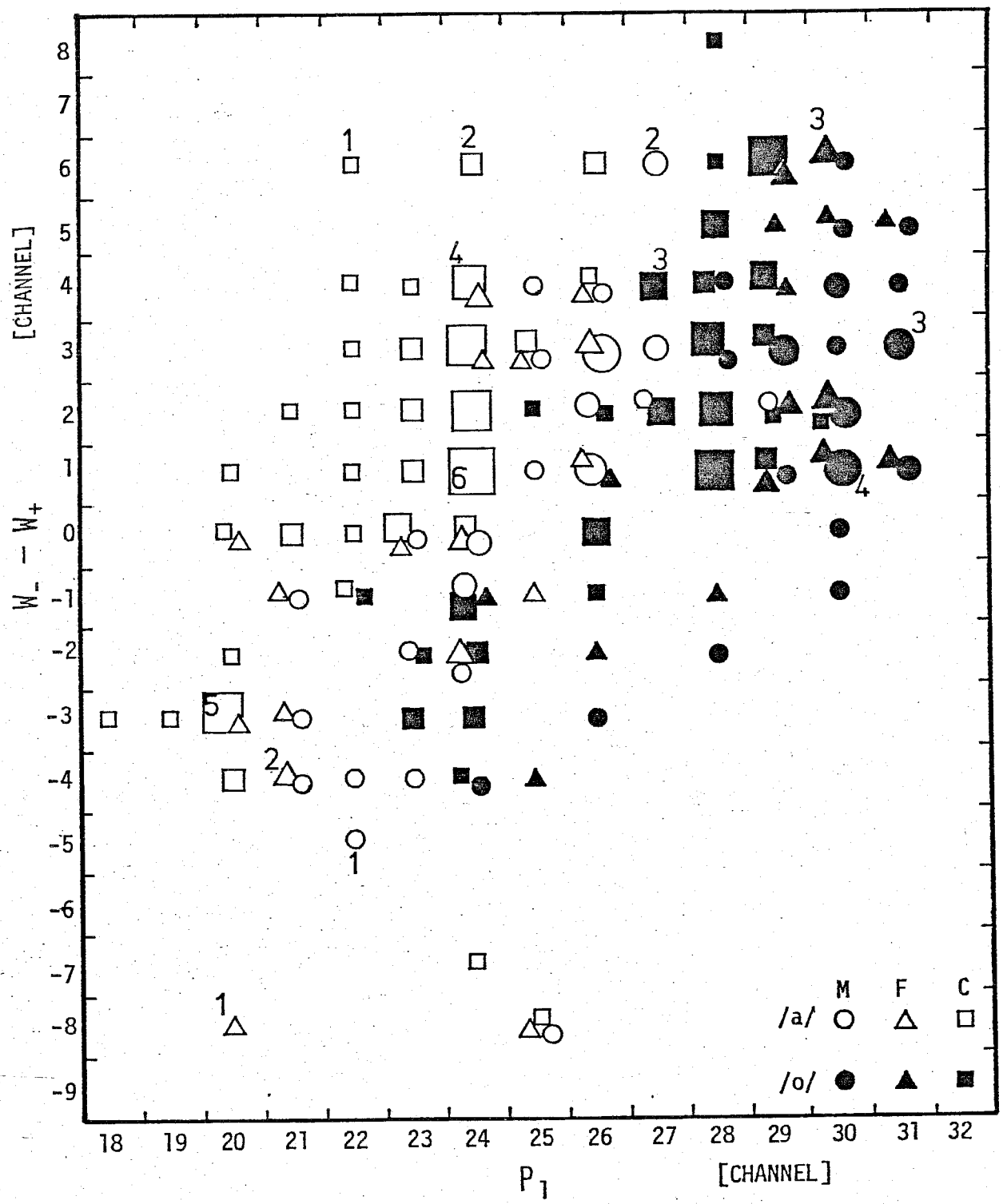


Fig. 4-11 /a/ と /o/ における P_1 と $(W_- - W_+)$ の関係.

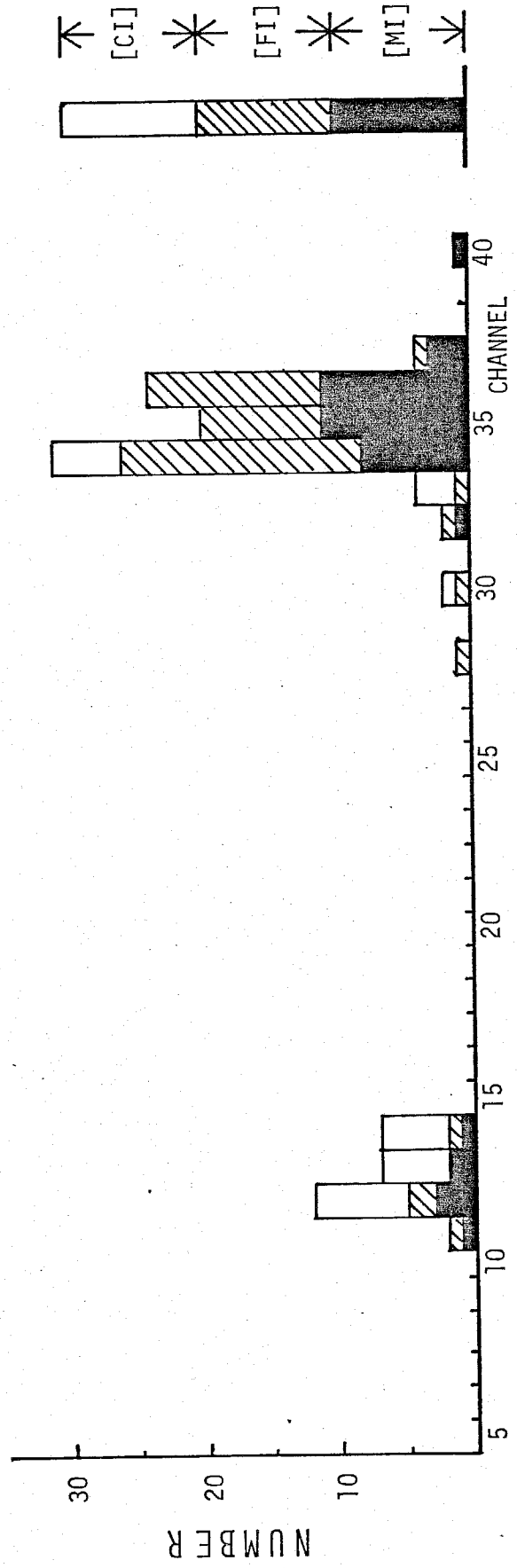


Fig. 4-12 /i/ における P1 (最大ピーク出現チャネル) の分布.

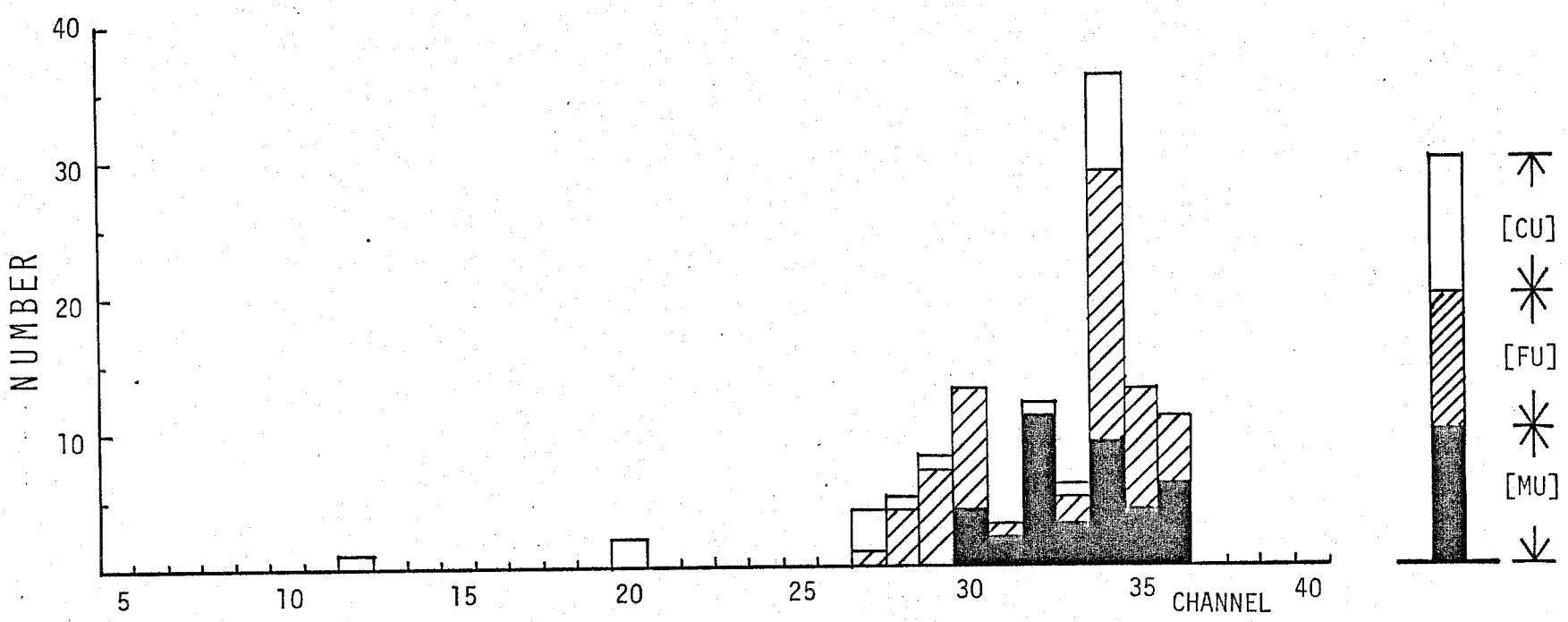


Fig. 4-13 /u/ における P_1 （最大ピーク出現チャネル）の分布.

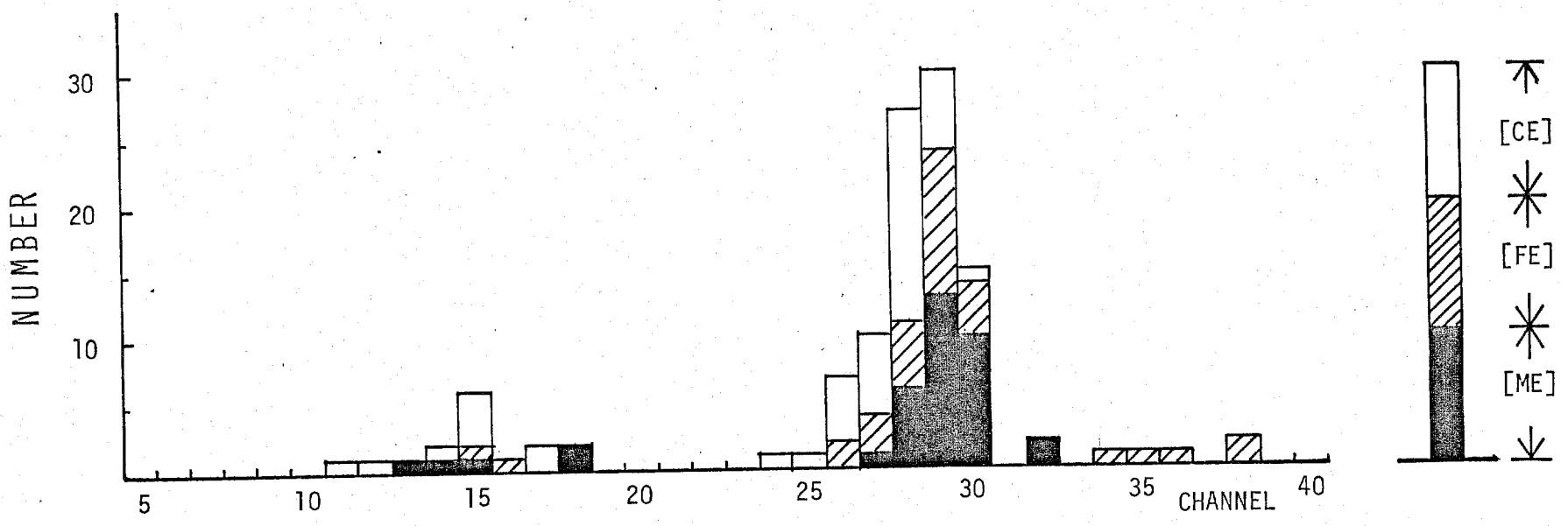
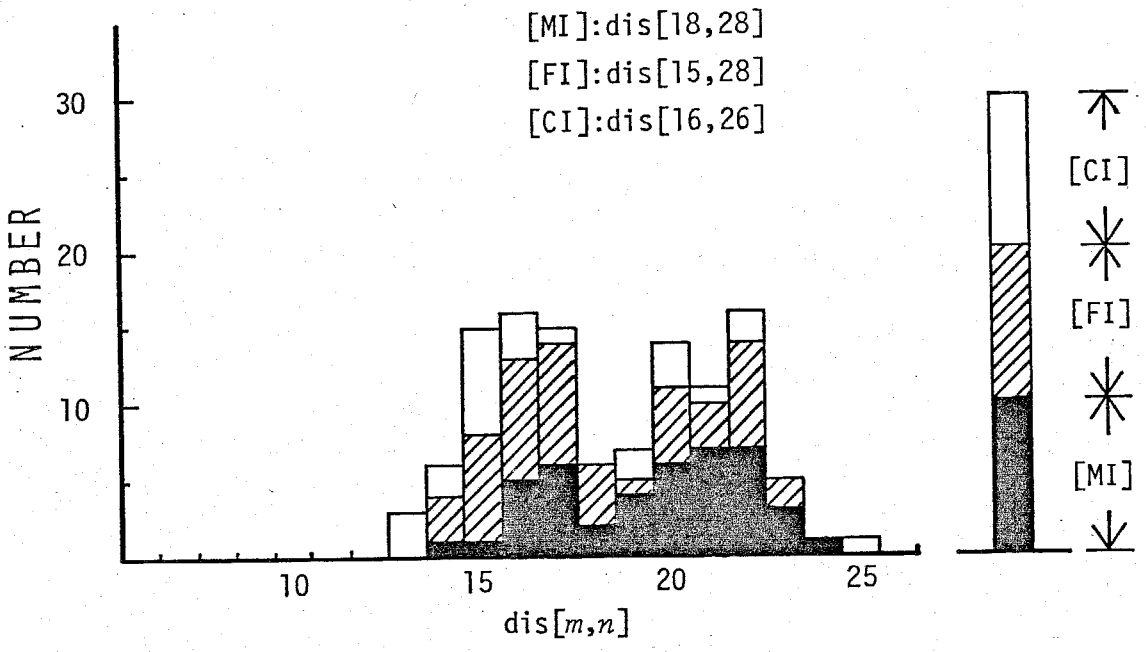
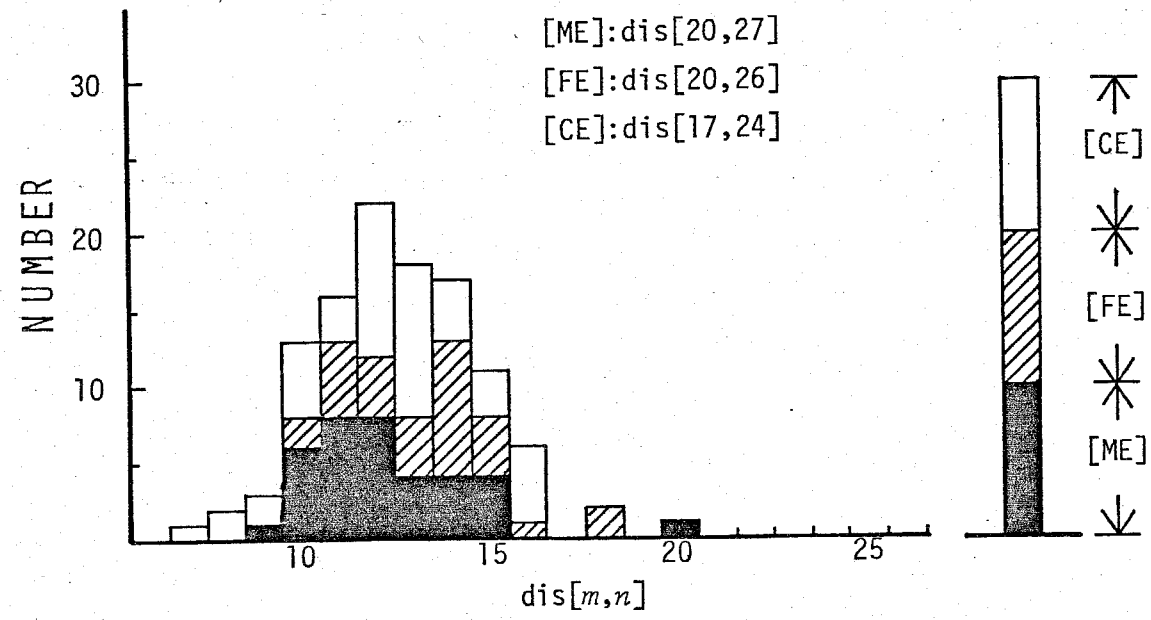


Fig. 4-14 /e/ における P_1 (最大ピーク出現チャネル) の分布.



Histograms of dis[m,n] for /i/.



Histograms of dis[m,n] for /e/.

Fig. 4-15 /i/ および /e/ における dis[m, n] のヒストグラム.

4.5 認識処理

Fig. 4-1に示したように，本母音認識システムは2段階の処理を行ない最終判断を下すシステムであるが，その各段階での処理について説明する。

(1). 第1段階.

入力音声のスペクトルパターンと15カテゴリ（[MA]～[CO]）のすべての参照パターンとの距離を計算する。入力スペクトルパターンを $X = (x_1, x_2, \cdots x_{54})$，あるカテゴリの参照パターンを $Y = (y_1, y_2 \cdots y_{54})$ とすると，X と Y との距離 l を（4-7）式で定義する。この式は，（4-4）式と同じものである。

$$\left.\begin{aligned} l &= \min\{l_i : i = 0, \pm 1\} \\ l_i &= \sum_{n=2}^{42} (\log y_n - \log x_{n+i})^2 \end{aligned}\right\} \quad (4-7)$$

距離 l_i は，参照パターン作成時と同様に入力スペクトルパターン X を ± 1 チャネルシフトしたものを含め計算し，その中の最小のものを入力スペクトルパターン X と参照パターン Y との距離 l とする。スペクトルパターン上のピークのような値の大きい部分のみならず，値の小さい箇所も含め形の違いを強調するために，距離は対数領域で計算する。

第1段階では，各参照パターンに対して l を求め，この値を使い第2段階へ移るための候補を上げる。以下，その手順を例を示しながら説明する。

まず距離 l の小さいカテゴリ順に並べ，閾値 θ 以下のカテゴリと θ を越える1カテゴリを選び，第2段階への候補とする。閾値 θ は，全サンプルについて l を計算した結果から，2.0と決定した。もし，第1段階で距離 l が θ 以下となるカテゴリが1つもない場合は，入力スペクトルパターンは母音のスペクトルパターンでないと判断され認識処理は終了する。

(例)

ℓ[MA] = 0.69
ℓ[FA] = 0.98
ℓ[MO] = 1.49
ℓ[CA] = 2.16
ℓ[FE] = 3.03
ℓ[FO] = 5.24
:

→第2段階へ
$\theta = 2.0$

(2). 第2段階.

第2段階では，第1段階で候補に上げられたカテゴリについて，距離ℓの小さい順に，入力スペクトルパターンがそのカテゴリに適するかどうかを，前節で示した4種の形態的特徴を表わすパラメータを用いて検証していく。検証の途中で合致するカテゴリがあれば，まだ未検証の候補カテゴリが残っていようとも，その時点で入力スペクトルパターンはそのカテゴリに属するものとして認識処理は終了する。この検証の結果，合致しない場合の状態としては「保留」と「除外」の2種をもうける。候補カテゴリの中で第2段階で合致するものがない場合は，「保留」と判断された候補カテゴリの中でℓの最小のカテゴリを認識結果とする。また検証結果のすべてが「除外」の場合は，入力スペクトルパターンは母音でないと判断する。

各カテゴリの検証アルゴリズムは，4種の特徴パラメータの中で各カテゴリ間の特徴の相違を良く表わしているものを選び，識別能力を上げるように作成した。Fig. 4-16 ~ Fig. 4-30に全カテゴリの検証アルゴリズムのフローチャートを示す。図中，○は「合致」，×は「除外」，△は「保留」を表わす。また，図中のP_n，D_n等の判定条件，例えば"$\exists n,\ 17 \leqq P_n \leqq 20$"は「17ch以後20ch以前にいずれかのピークが存在するか？」という意味を表わす。なおこの検証アルゴリズムの規模はFORTRAN言語で約400 stepである。

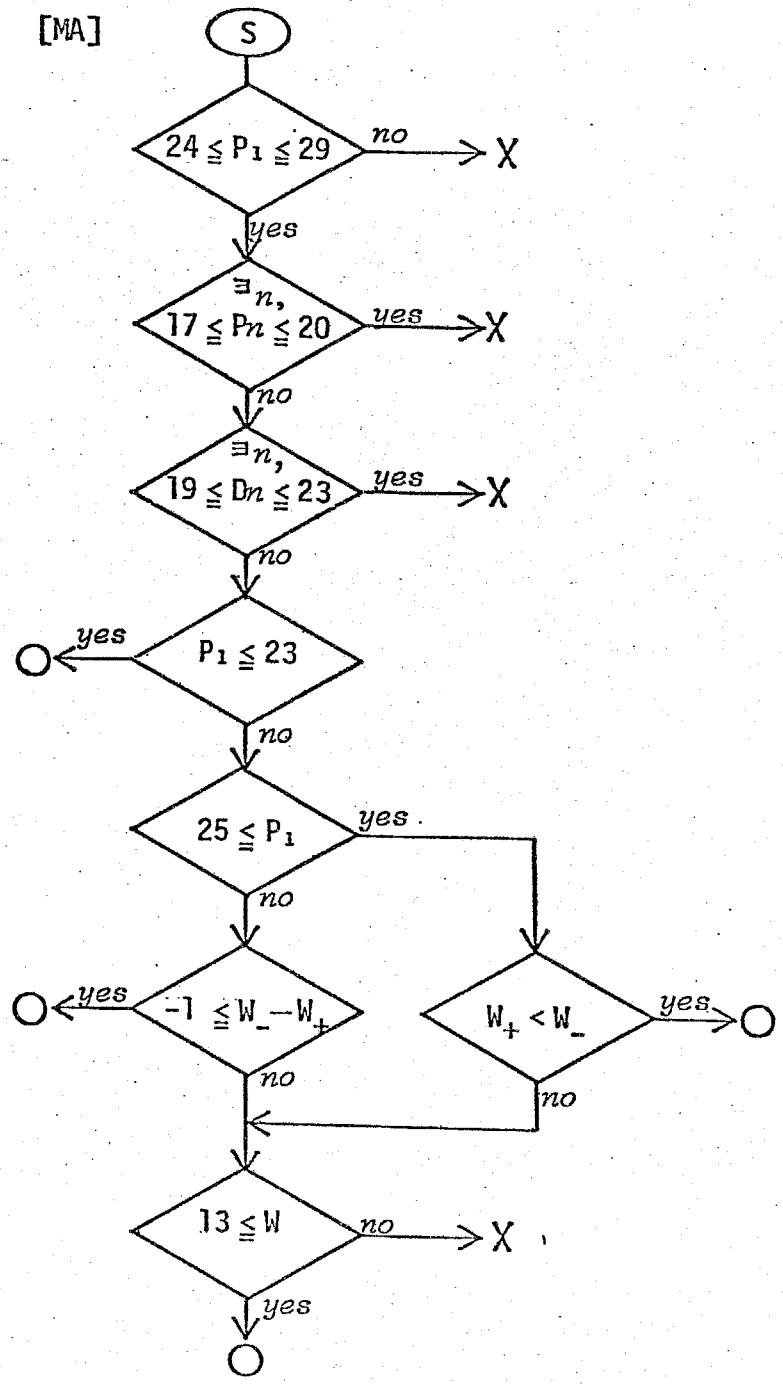


Fig. 4-16 [MA]の検証アルゴリズム.

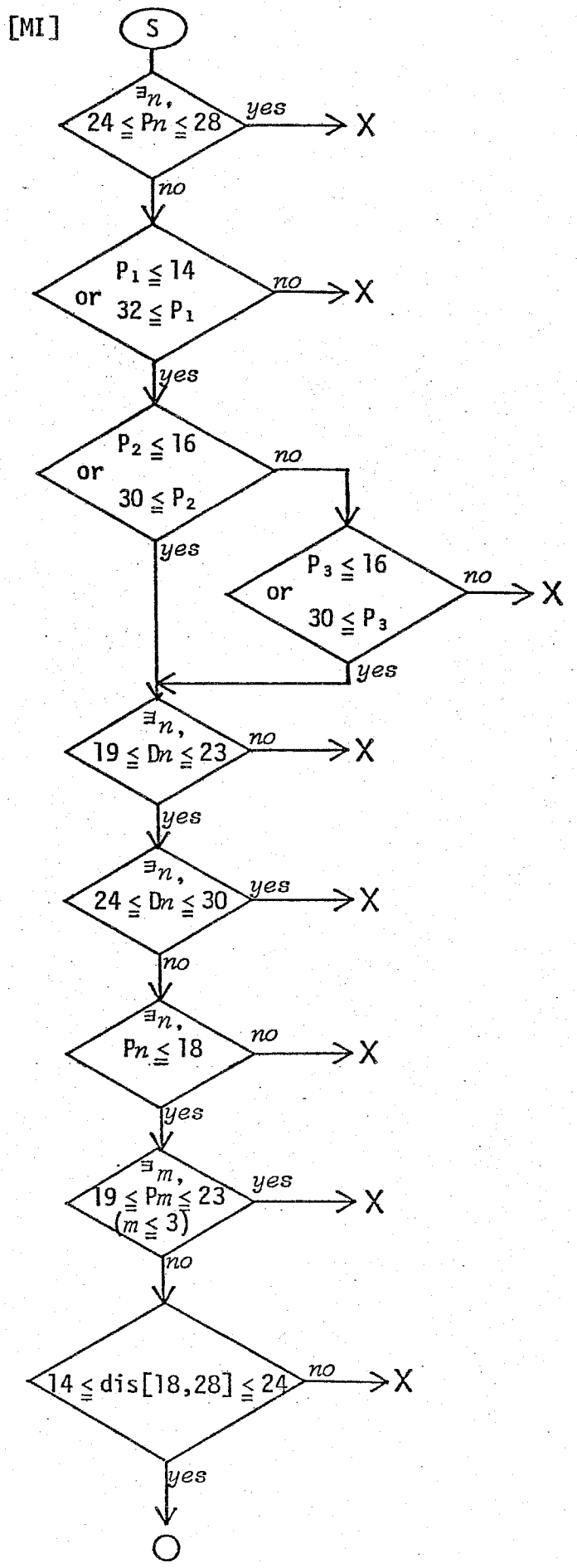


Fig. 4-17 [MI]の検証アルゴリズム.

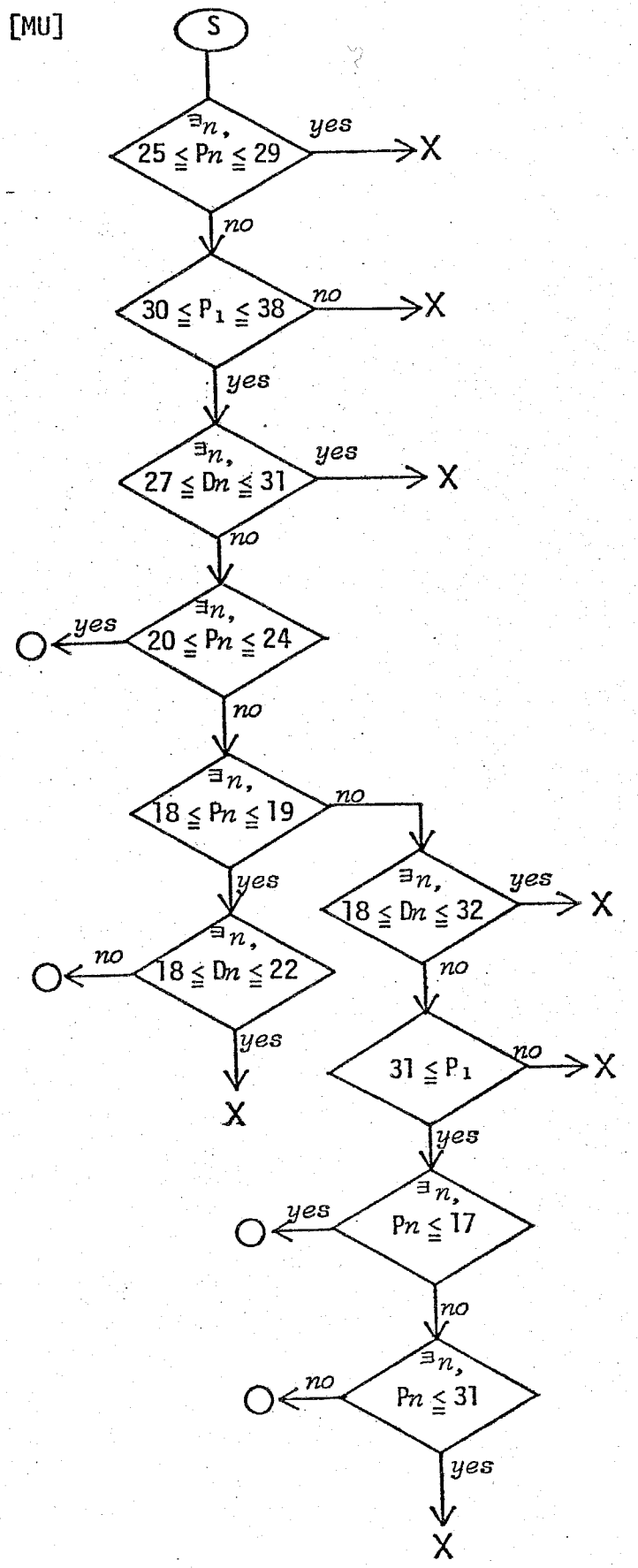


Fig. 4-18 [MU]の検証アルゴリズム.

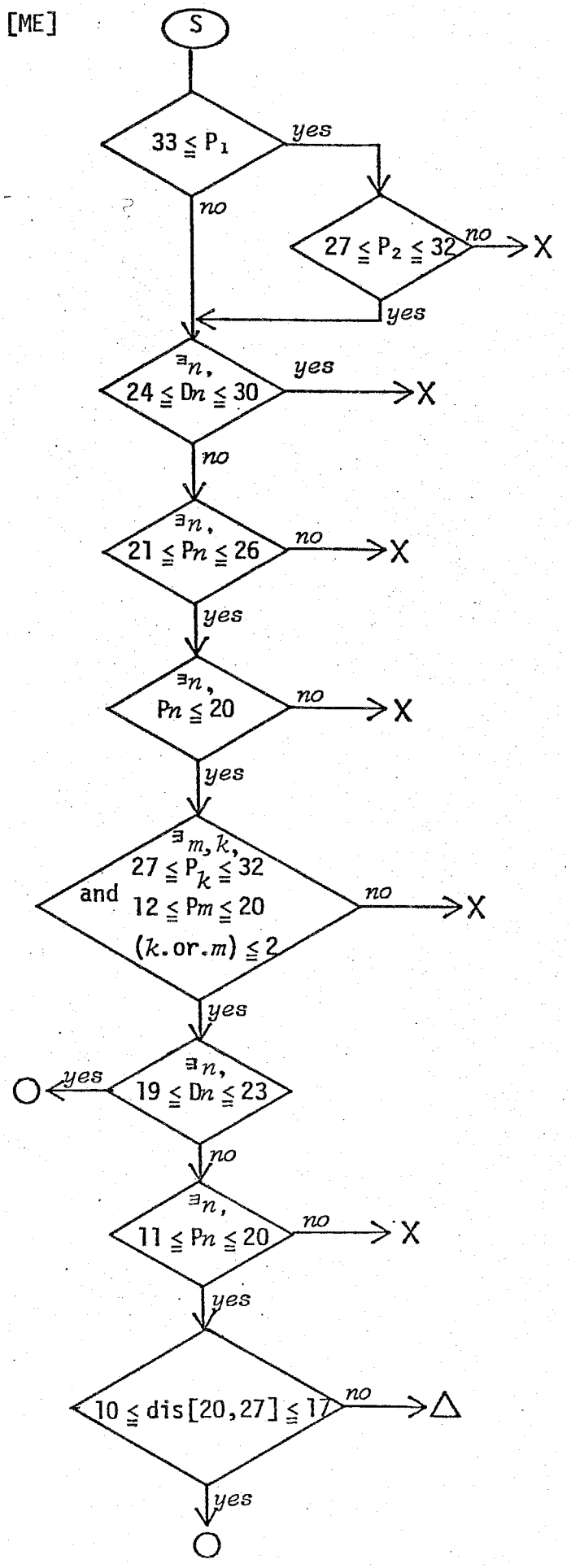


Fig. 4-19 [ME] の検証アルゴリズム.

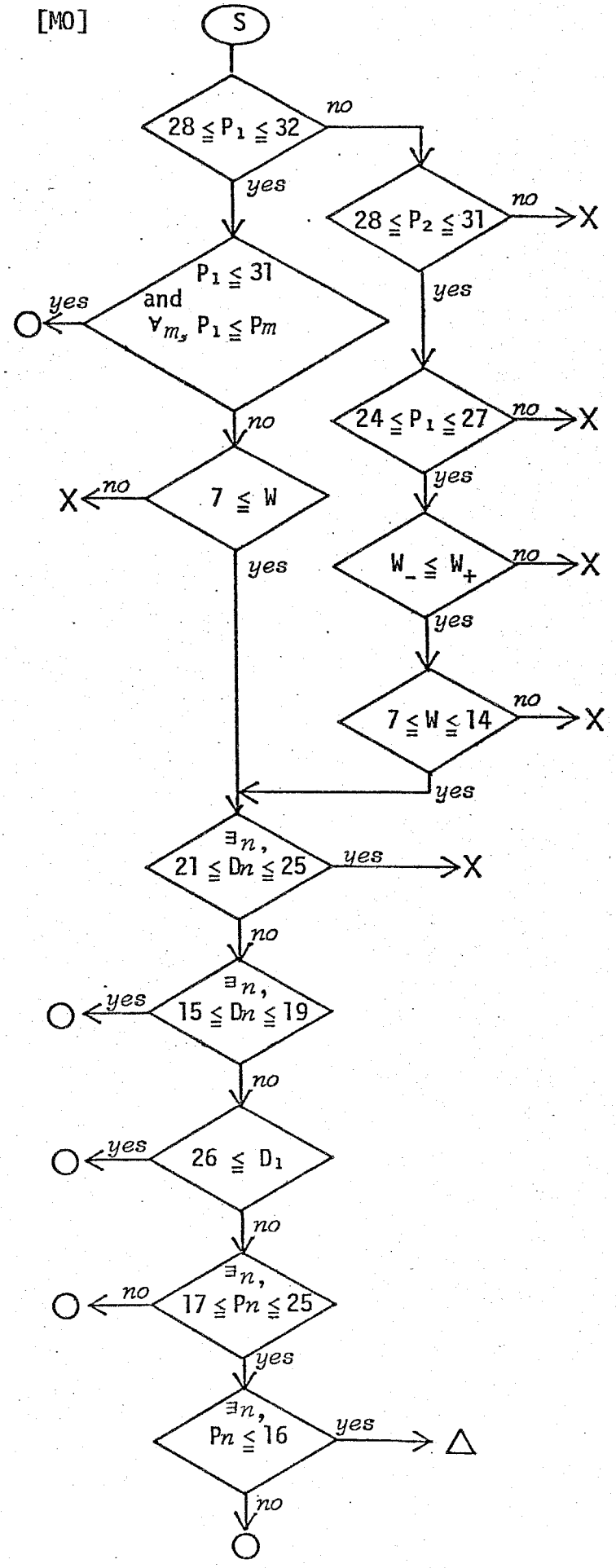


Fig. 4-20 [MO]の検証アルゴリズム.

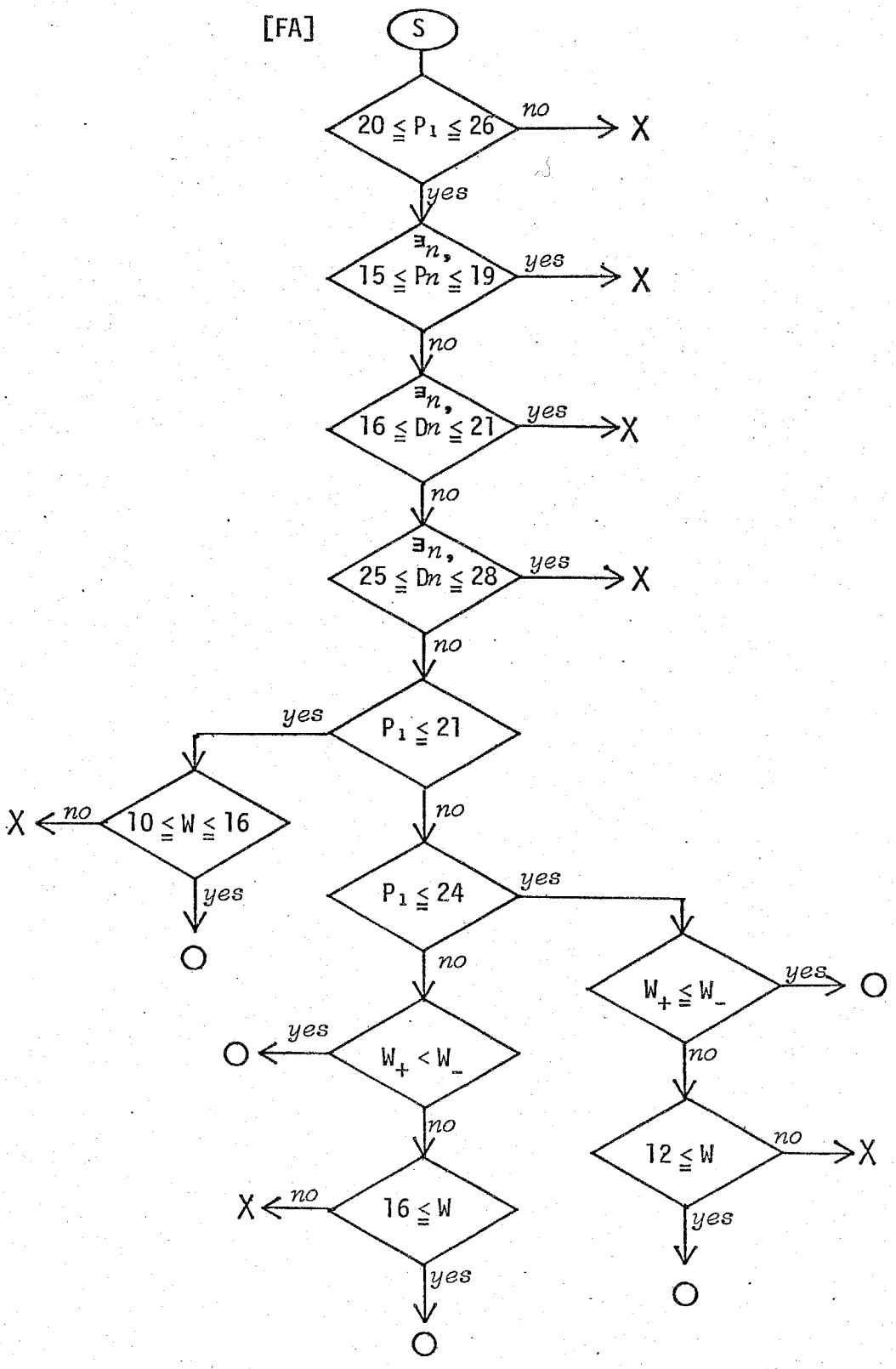


Fig. 4-21 [FA] の検証アルゴリズム.

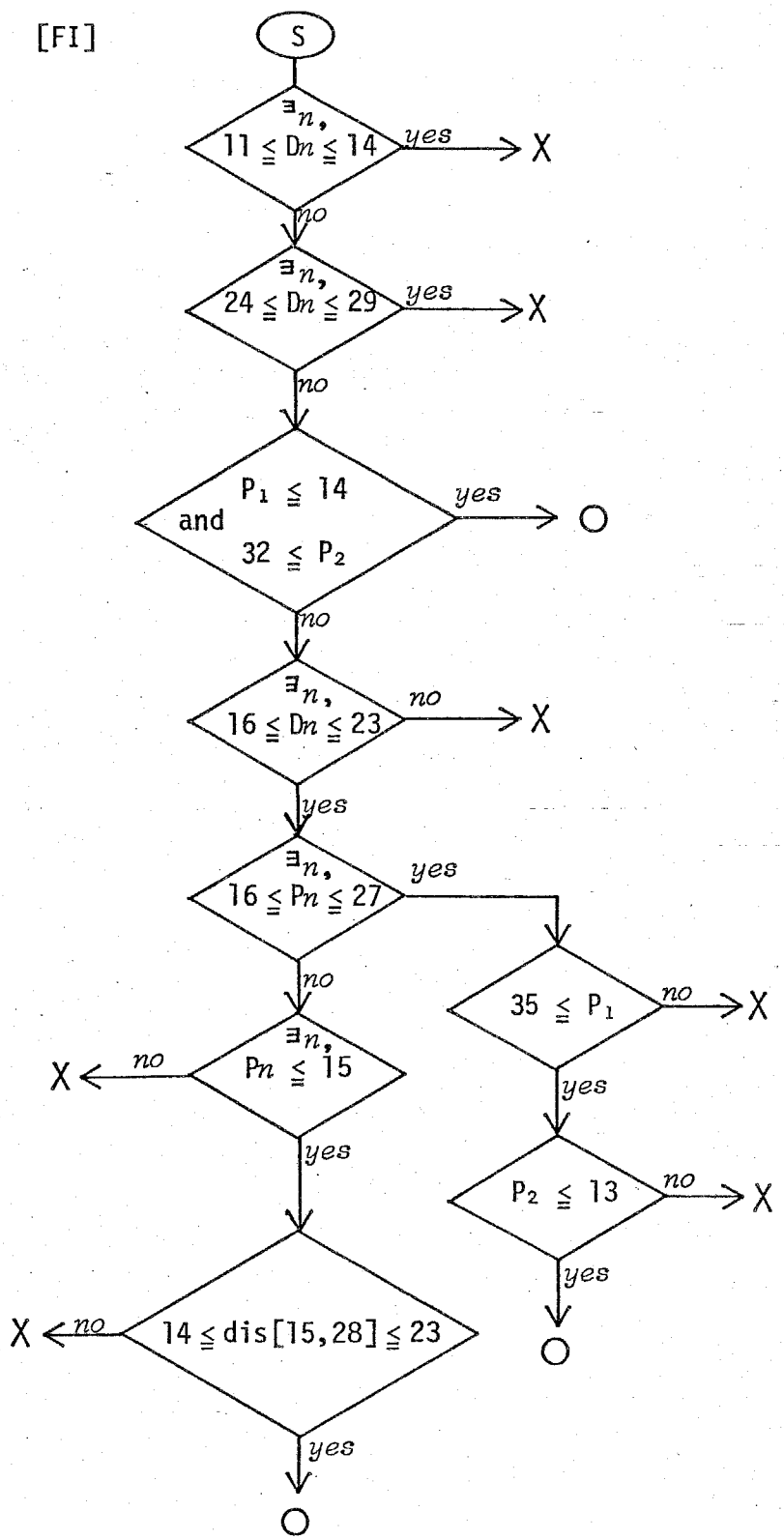


Fig. 4-22 [FI] の検証アルゴリズム.

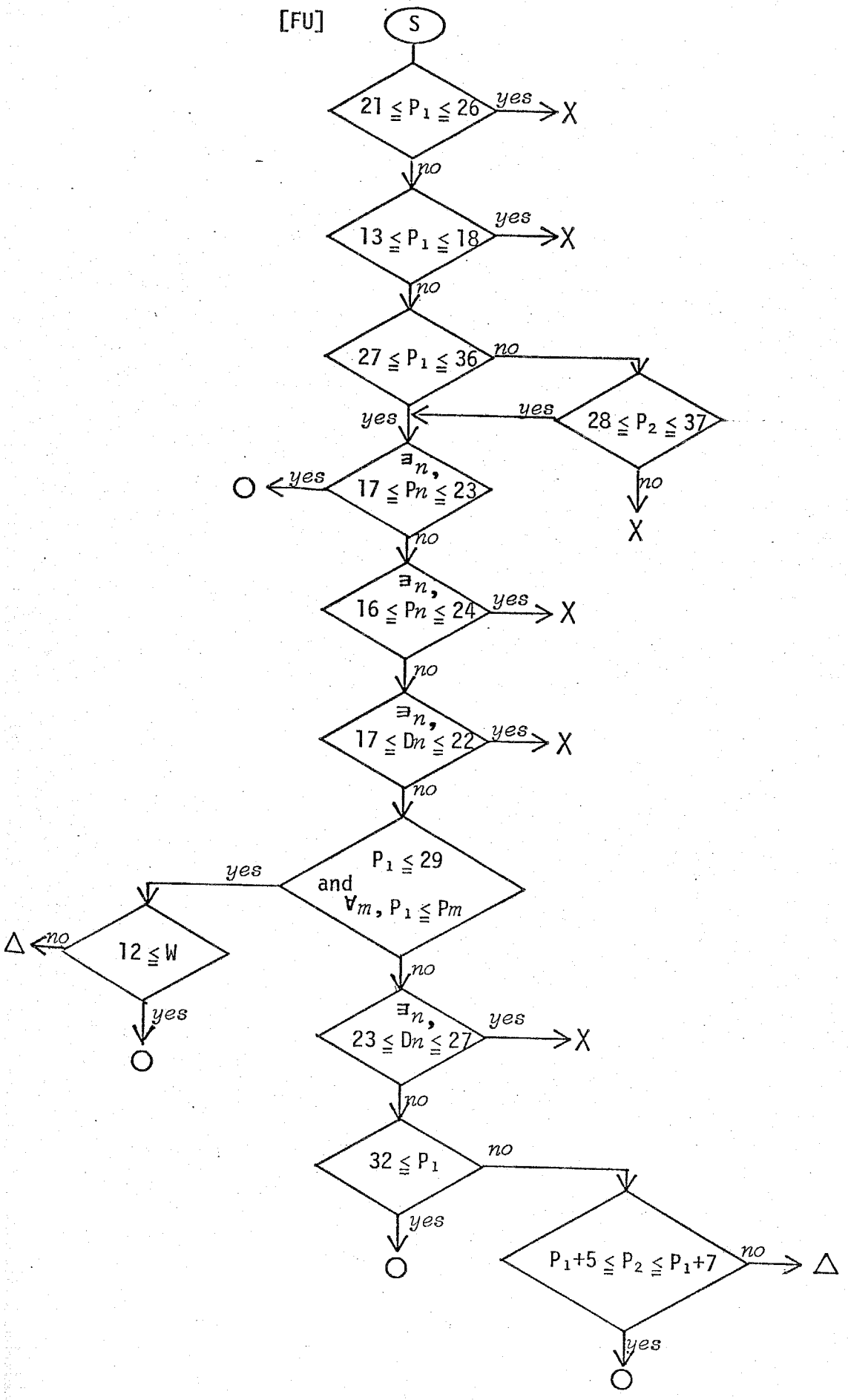


Fig. 4-23 [FU] の検証アルゴリズム.

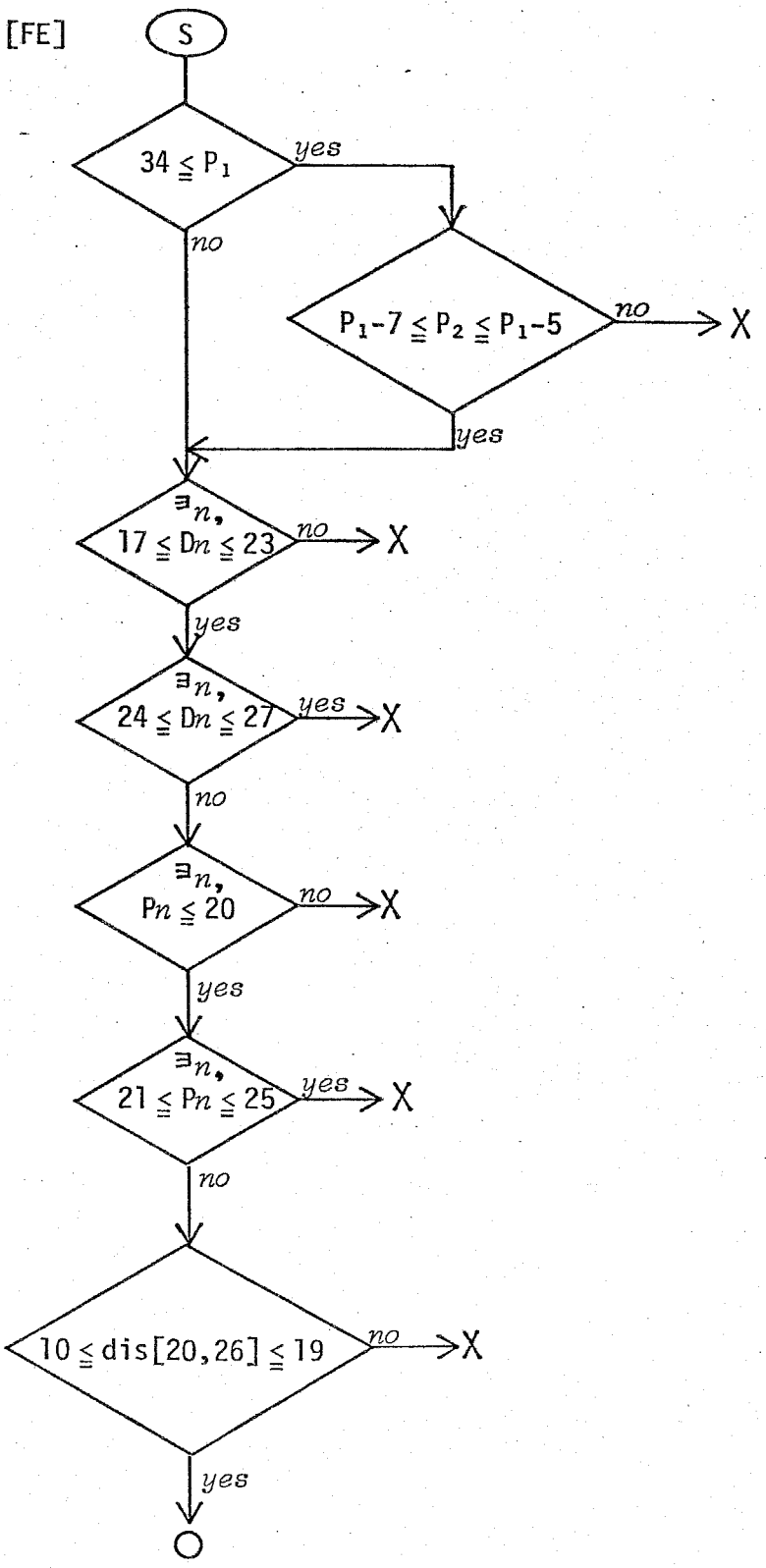


Fig. 4-24 [FE] の検証アルゴリズム.

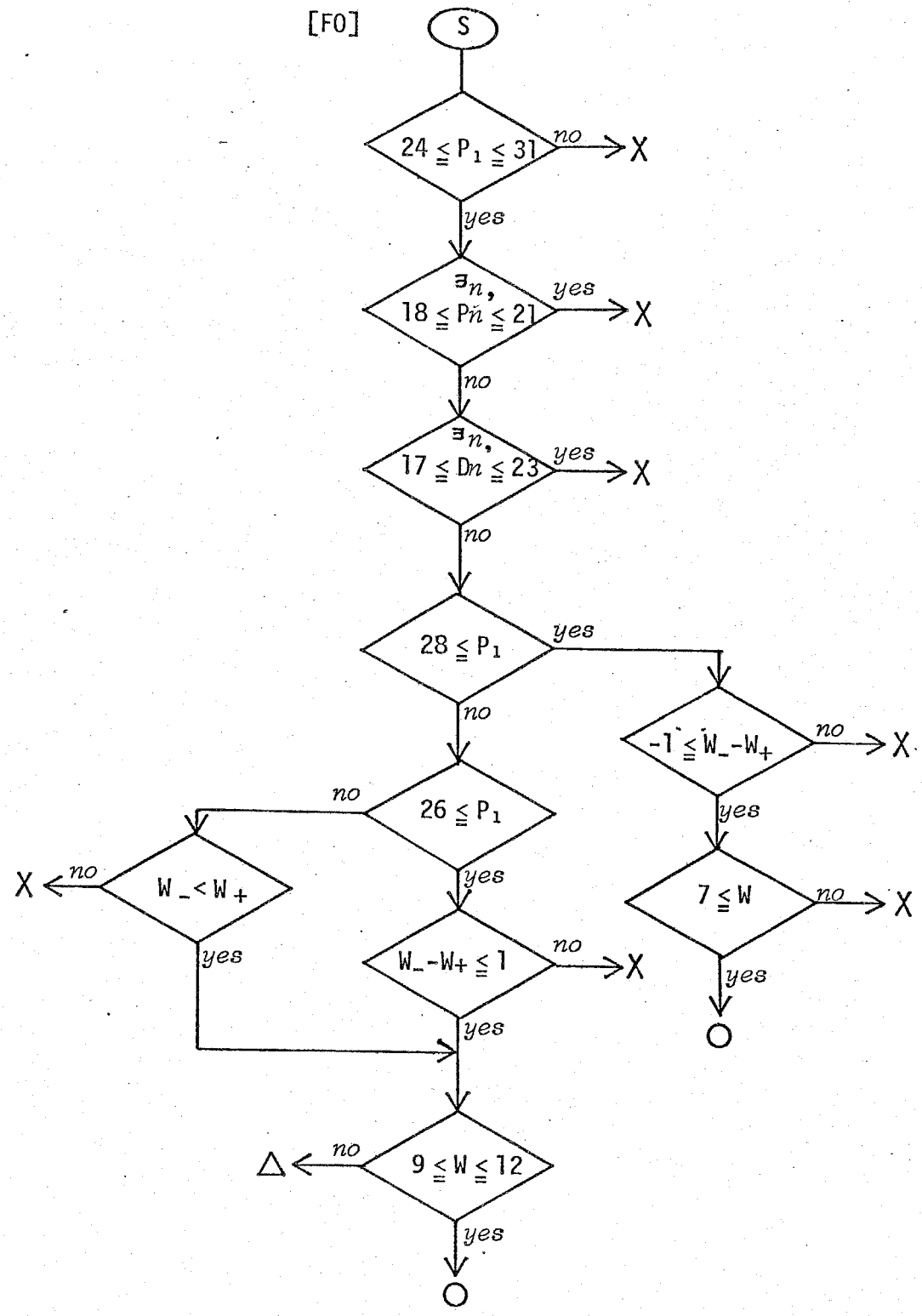


Fig. 4-25 [F0] の検証アルゴリズム.

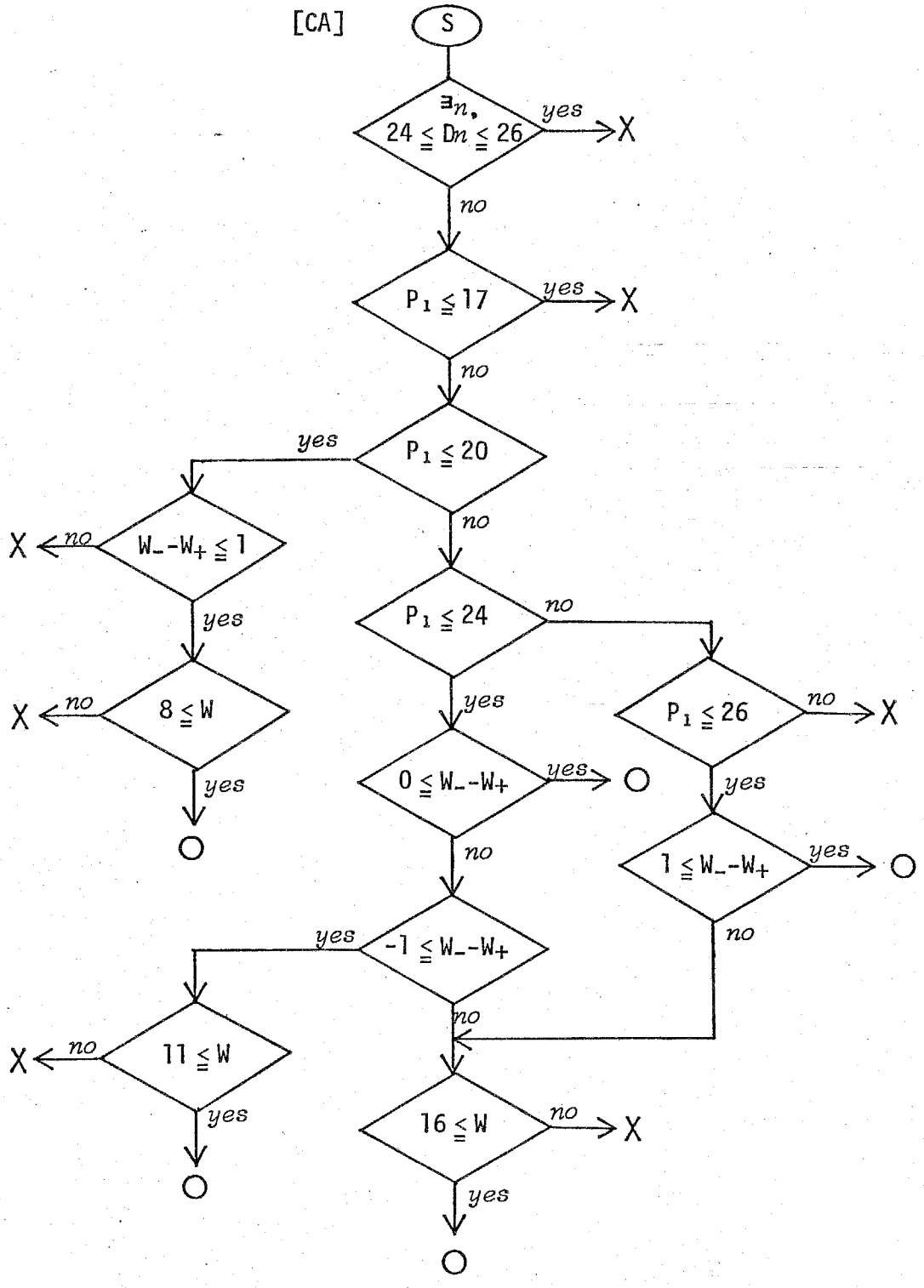


Fig. 4-26 [CA] の検証アルゴリズム.

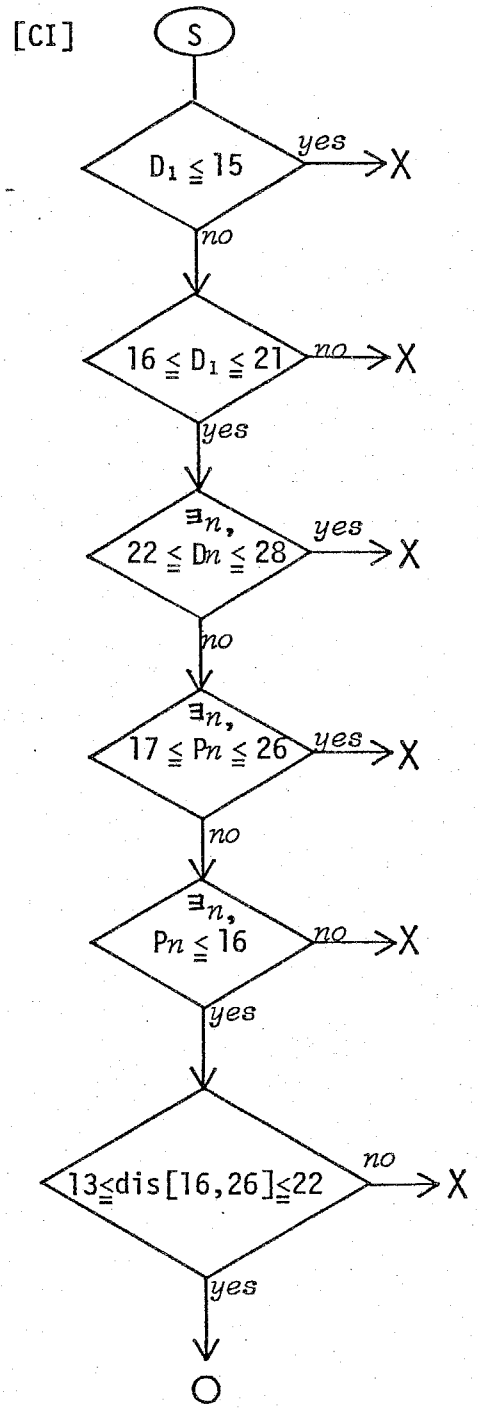


Fig. 4-27 [CI]の検証アルゴリズム.

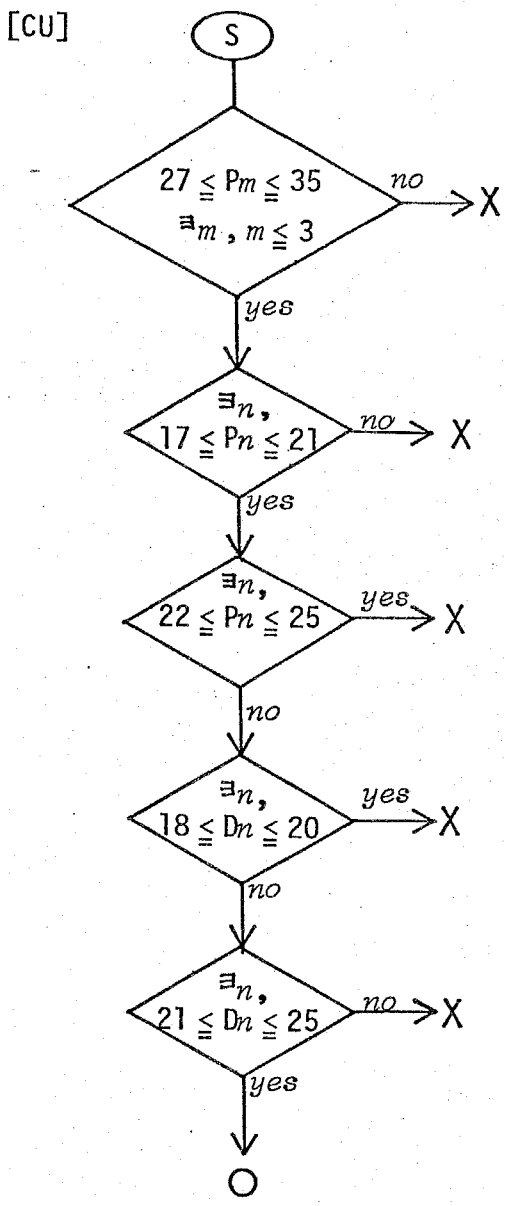


Fig. 4-28 [CU] の検証アルゴリズム.

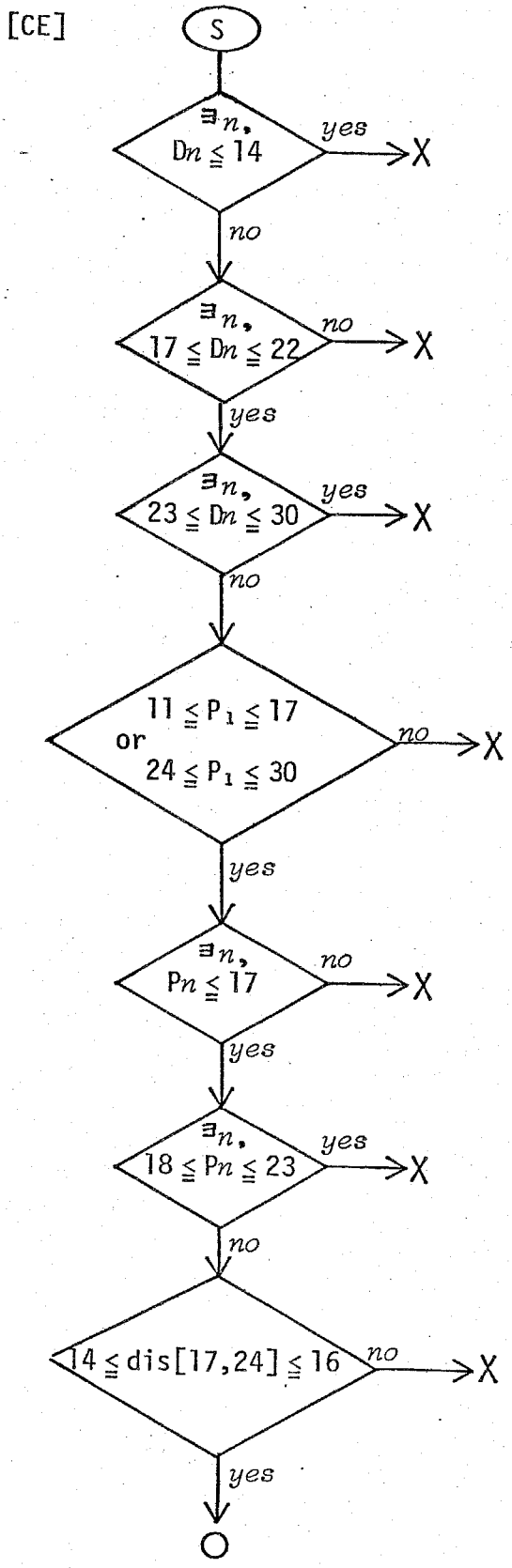


Fig. 4-29 [CE]の検証アルゴリズム.

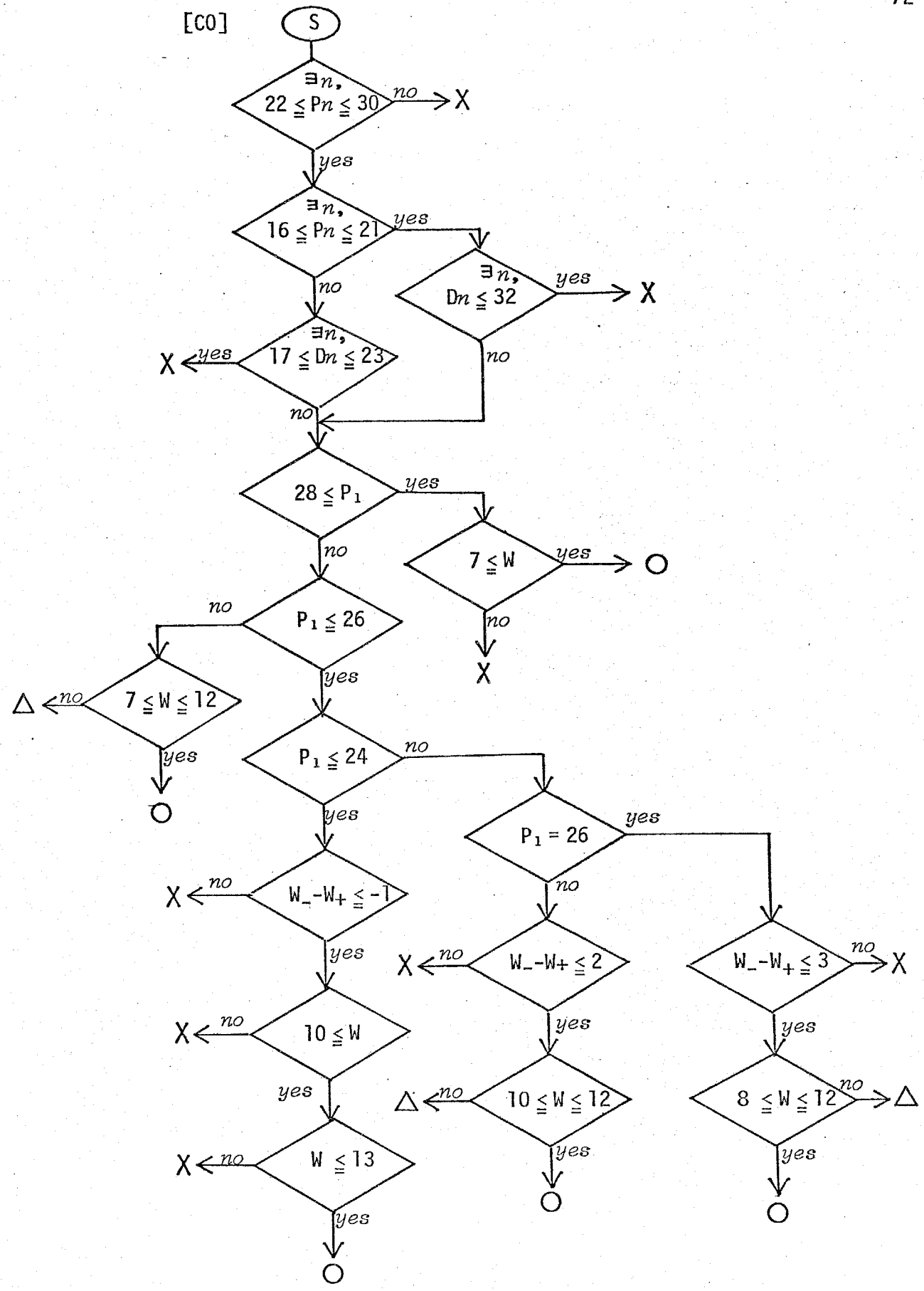


Fig. 4-30 [C0]の検証アルゴリズム.

4.6 認識実験.

認識システムを構成する際に用いたと同じ音声（成人男性32名，同女性25名，10～12才男子44名，同女子16名の発声した単母音，合計585個）について認識実験を行なった。フレーム周期12.8 msec として連続する8フレーム（1フレーム12.8 msec 計102.4 msec）のスペクトルパターンの認識結果から，音韻クラスに関する次のような多数決をとり最終認識結果とした。

①. 最初の1フレームを除く7フレームの認識結果について，音韻クラス間の多数決をとる。

②. ①において，2つ以上の音韻クラスが同票であった場合は，最初の1フレームの認識結果を加え多数決をとる。

以下に，認識例を3例示す。また．Table 4－1に 最終結果のConfusion Matrixを示すが，成人男性で100%，同女性で96.8%，子供で98.0%，総合で98.3%という高い認識率を得ることができた。

ex. 1)　INPUT VOICE (male adult /a/)　RESULT (/a/)

TIME	RESULT	NN	1st. STEP	P1	P2	P3	P4	P5	P6	D1	D2	D3	D4	W_-	W_+	W
1	MA	5	MA FA MO ME FE FO MU CA CO FU FI MI	27	22	34	39	42	46	24	33	37	0	9	4	13
2	MA	4	MA MO FA ME MU FO FE MI FU CO CA FI	27	22	41	35	0	0	32	37	0	0	9	3	12
3	MA	3	MA MO FA ME MU FO FE MI FU CO FI CA	27	22	40	35	48	0	31	37	47	0	8	3	11
4	MA	4	MA MO FA ME MU FO FE MI FU CO CA FI	27	22	41	35	0	0	32	37	48	50	8	3	11
5	MA	4	MA MO FA ME MU FO FE MI FU CO FI CA	27	22	40	34	48	0	32	37	49	0	8	3	11
6	MA	4	MA MO FA ME MU FO FE MI FU CO CA FI	27	22	41	34	47	0	32	36	0	0	8	3	11
7	MA	4	MA FA MO ME MU FO FE MI FU CO CA FI	27	41	35	48	0	0	32	37	0	0	8	3	11
8	MA	4	MA FA MO ME MU FO FE MI FU CO CA FI	26	22	40	35	0	0	31	37	50	0	7	4	11

第1段階の結果（左から L の小さい順）

↑ 候補カテゴリ数

ex. 2)　INPUT VOICE (female adult /e/)　RESULT (/e/)

TIME	RESULT	NN	1st. STEP	P1	P2	P3	P4	P5	P6	D1	D2	D3	D4	W_-	W_+	W
1	ME	7	ME MU MO FE MA FA FO FU FI MI CO CA	36	30	38	18	41	45	33	0	0	0	14	4	18
2	MU	2	MU ME FO MO FU MA FI MI FE FA CO CA	38	32	45	48	0	0	34	0	0	0	8	3	11
3	???	2	MU FO ME MO FU MA FE FI FA MI CO CA	38	32	30	47	0	0	33	44	0	0	5	3	8
4	ME	4	MU ME FO MO FU FE MA FI FA MI CO CA	38	32	28	18	47	0	19	34	45	0	8	3	11
5	ME	3	MU ME FO MO FU FE FI MA MI FA CO CA	38	32	18	47	0	0	19	33	49	0	8	3	11
6	MU	2	MU ME FO MO FI FU MI FE MA FA CO CA	38	32	18	45	48	0	34	47	49	0	8	3	11
7	ME	3	MU ME FO MO FE FU FI MA MI FA CO CA	38	32	18	45	0	0	19	34	48	0	9	3	12
8	MU	3	MU ME MO FO FE MA MI FI FU FA CO CA	38	32	24	17	46	0	20	34	48	0	9	3	12

ex. 3)　INPUT VOICE (female child /a/)　RESULT (/a/)

TIME	RESULT	NN	1st. STEP	P1	P2	P3	P4	P5	P6	D1	D2	D3	D4	W_-	W_+	W
1	CA	5	CA FA FE MA CE CO FO MO CU FU ME FI	24	18	35	47	0	0	31	0	0	0	8	6	14
2	CA	7	CA FE FA CO FO FU CE CU MA MO ME FI	25	35	18	42	0	0	31	41	0	0	7	5	12
3	CA	6	FE FA CO CA FU CE FO CU MA MO ME FI	25	35	18	46	48	44	31	43	45	47	8	5	13
4	CA	7	FE FA CO CE FU CA FO CU MA MO ME FI	24	35	18	47	0	0	32	45	50	0	5	6	11
5	CA	6	FE FA CA CE CO FU CU FO MA MO ME FI	25	35	18	43	45	48	31	47	0	0	8	5	13
6	CA	8	FE FA CO FU FO CE CA CU MA MO ME FI	24	35	18	44	0	0	32	0	0	0	5	6	11
7	CA	6	FE FA CA CE CU CO FU FO MA MO ME FI	25	35	18	45	0	0	31	0	0	0	8	5	13
8	CA	6	FE FA CO FU FO CA CE CU MA MO ME FI	25	35	18	45	0	0	32	0	0	0	6	5	11

Table 4-1 単母音認識実験結果の Confusion Matrix.

INPUT VOWEL			RESULT /a/	/i/	/u/	/e/	/o/	SCORE (%)
	MALE	/a/	32					100.0 %
		/i/		32				
		/u/			32			
		/e/				32		
		/o/					32	
	FEMALE	/a/	25					96.8 %
		/i/		25				
		/u/			24	1		
		/e/	1			23	1	
		/o/			1		24	
	CHILD	/a/	59				1	98.0 %
		/i/		59		1		
		/u/			58	1	$1_{(1)}$	
		/e/				58	$2_{(1)}$	
		/o/					60	

( ):female child

4.7 検討

本単母音認識システムが成功した理由について検討してみると，次の3点が上げられよう。

(1). 参照パターンとの距離を利用する方法と，スペクトルパターンの形態的特徴を利用する方法の2つの識別法が相補的に組み合わされている。

認識の第1段階で最小距離となったカテゴリを認識結果とし，8フレーム間で音韻クラス間の多数決により最終決定を下した場合の結果を Table 4-2 に示すが，成人男性で89.4%，同女性で90.4%，子供で85.3%，全体で87.5%の識別率しか得られていない。また，第2段階のみの識別能力については，入力音声のスペクトルパターンの形態的特徴が合致したすべてのカテゴリについて，8フレーム間の多数決により決定したが，全体で79.8%の識別率しか得られていない。これらの結果から，簡単な構成で単独ではあまり高い識別能力を持たない識別法を組み合わせることにより，高い識別能力を持つ認識システムが構成できることが示された。

(2). 男性，女性，子供の音声をすべてまとめて取り扱い，話者の属性決定を音韻性決定以前に独立して行なうシステムを採用していない。

一般に行なわれている音声認識の研究では，まず取り扱う話者の属性（性別・年令）を限定し，それらの属性ごとに音声の特徴をつかみ認識システムを構成したり，[31] 参照パターンの補正を行なったりする[32] のが普通である。しかし，それらの方法で不特定話者の音声認識システムを構成する場合，音韻認識以前に話者の属性を決定する必要が生じ，その話者の属性決定段階での判定結果が音韻認識に悪影響を及ぼすことが考えられる。極端な例として，成人男性がわざとピッチを高めて発声した場合を考えると，その声はピッチは高くともフォルマント周波数はほとんど変化しないので，ピッチを用いて属性を決定してしまうシステムでは誤認識を生じることになろう。本システムでは，

第1段階で参照パターンとの距離の算出の時点でスペクトルパターンのピッチ周波数を含むチャネルまでの値が使われることにより，ピッチの違いが距離の大小に影響するが，その大小の差がピッチの違いによるものかフォルマント構造の違いによるものかは判断されず，同一音韻クラスの候補（4．5節 p.57 の例では［MA］，［FA］，［CA］の3候補）がいくつも上げられ第2段階での判断にゆだねられるので，話者の属性を前もって決定してしまうことによる誤りを減らすことができる。

Table 4-2 第1段階の最小距離カテゴリを認識結果とした場合の Confusion Matrix.

			RESULT					SCORE (%)
			/a/	/i/	/u/	/e/	/o/	
INPUT VOWEL	MALE	/a/	31		1			89.4 %
		/i/		30	2			
		/u/			31	1		
		/e/	3		3	22	4	
		/o/	1		2		29	
	FEMALE	/a/	25					90.4 %
		/i/		23	2			
		/u/	1		20	1	3	
		/e/	1		2	20	2	
		/o/					25	
	CHILD	/a/	51			3	6	85.3 %
		/i/		58	1	1		
		/u/		2	51	1	6	
		/e/			11	44	5	
		/o/	1		1	6	52	

(3). 基底膜モデルのQ値が約2と低いために，話者の個人差がスペクトルパターン上に顕著に現われない。

このために特に参照パターンとの距離計算において，入力スペクトルパターンの大局的な特徴が良くとらえられるものと考えられる。また，第2段階の検証アルゴリズムの構造も簡単なものにすることができたと考えられる。

今回の認識実験で誤認識した音声のうち，4例の/e/はスペクトルパターンがFig. 4-2に見られるような2峰性の形状を示さなかったものである。試聴の結果，これらの音声はすべて正しく聴き取ることができたので，誤認識の原因は基底膜モデルのQ値の低さにあると考えられる。基底膜モデルのQ値をいくらに設定すればよいかという問題は，話者の個人差による変動をできるだけ小さくするという要請との兼ね合いで難しい問題であるが，新たな生理学上の成果をも考慮し検討を加えていく必要があろう。

他の誤認識例6例について見てみると，第1段階での結果が影響しているものが3例，第2段階の検証アルゴリズムにより誤認識されているものが2例である。第1段階で用いる参照パターンの作成に効果的な学習法を導入したり，第2段階の検証アルゴリズムの作成法にさらに検討を加えていくことにより改善は可能であろう。

4.8 むすび

基底膜モデルから得られるスペクトルパターンを用いて簡単な構造の不特定話者対応単母音認識システムを構成し，成人男性32名，同女性25名，10～12才の男子44名，同女子16名の発声した合計585個の単母音に対して認識実験を行なった結果，成人男性については100%，同女性96.8%子供98.0%，総合で98.3%という高い認識率を得ることができた。実験に用いた音声データは，すべての年令層をカバーしてはいないが，多数の変声期前の子供も含まれていることから，不特定話者の音声に十分対応できるシステムであるといえよう。

本システムの処理時間は，音声1フレーム（12.8 msec）あたり，基底膜演算に7 sec，認識処理に1 sec 合計8 secを要するが，この処理時間の短縮化の問題は本システムの応用の問題も含め次章で論ずる。

第5章　連続音声認識への応用

5.1　まえがき

第4章で示した単母音認識システムを応用して，不特定話者対応の連続音声認識システムを構成し，この母音認識システムが連続音声に対しても有効であることを検証する実験を行なった。子音に関しては，ただ子音部と母音部を区別するだけにとどまらず，将来の応用性も考慮し，4つの大分類クラスと補足的な4クラスに分類することを試みた。4つの大分類クラスとは，無声破裂音（p，t，k），無声摩擦音（s，sh，ts，ch，h），有声摩擦音（z），その他の有声子音（N，m，n，r，b，d，g，w，y）とし，補足的な4クラスとは，k，h，破擦音（ts，ch），及び撥音（N）である。連続音声データとしては，NHKのニュース放送を用いた。また，この連続音声認識システムでの認識処理の基本単位は，単母音認識の場合と同様に，フレーム長12.8 msec フレーム周期12.8 msec で出力される基底膜スペクトルパターンである。Fig. 5-1にこの連続音声認識システムの概略図を示す。同図中「PWR」は，基底膜各チャネルのパワーの総和を示し，「HP」は高周波帯域7～124チャネル（CF7，1.3 kHz ～ 4.0 kHz）へのパワーの集中度を，「BP」はピッチ周辺のパワーの集中度をそれぞれ百分率で表わしたものである。「PWR」は無音部や子音検出のためのセグメンテーション，母音区間決定のためのサブセグメンテーションに使われ，「BP」，「HP」は子音の検出・分類に使われる。

男性アナウンサー8名，女性アナウンサー5名の連続音声データに対して，この連続音声認識システムを適用したところ，母音認識率が80.0%（男性85.7%，女性72.9%），子音の識別率が72.7%という値が得られた。この母音認識率は，従来の不特定話者の単語音声認識システムや連続音声認識システムの母音認識率[(1)(2)]と比べて非常に高い値であり，この母音認識システムの有効性が確かめられた。

以下，この連続音声認識システムについて，セグメンテーション，母音認識，子音の検出・分類について詳しく述べる。

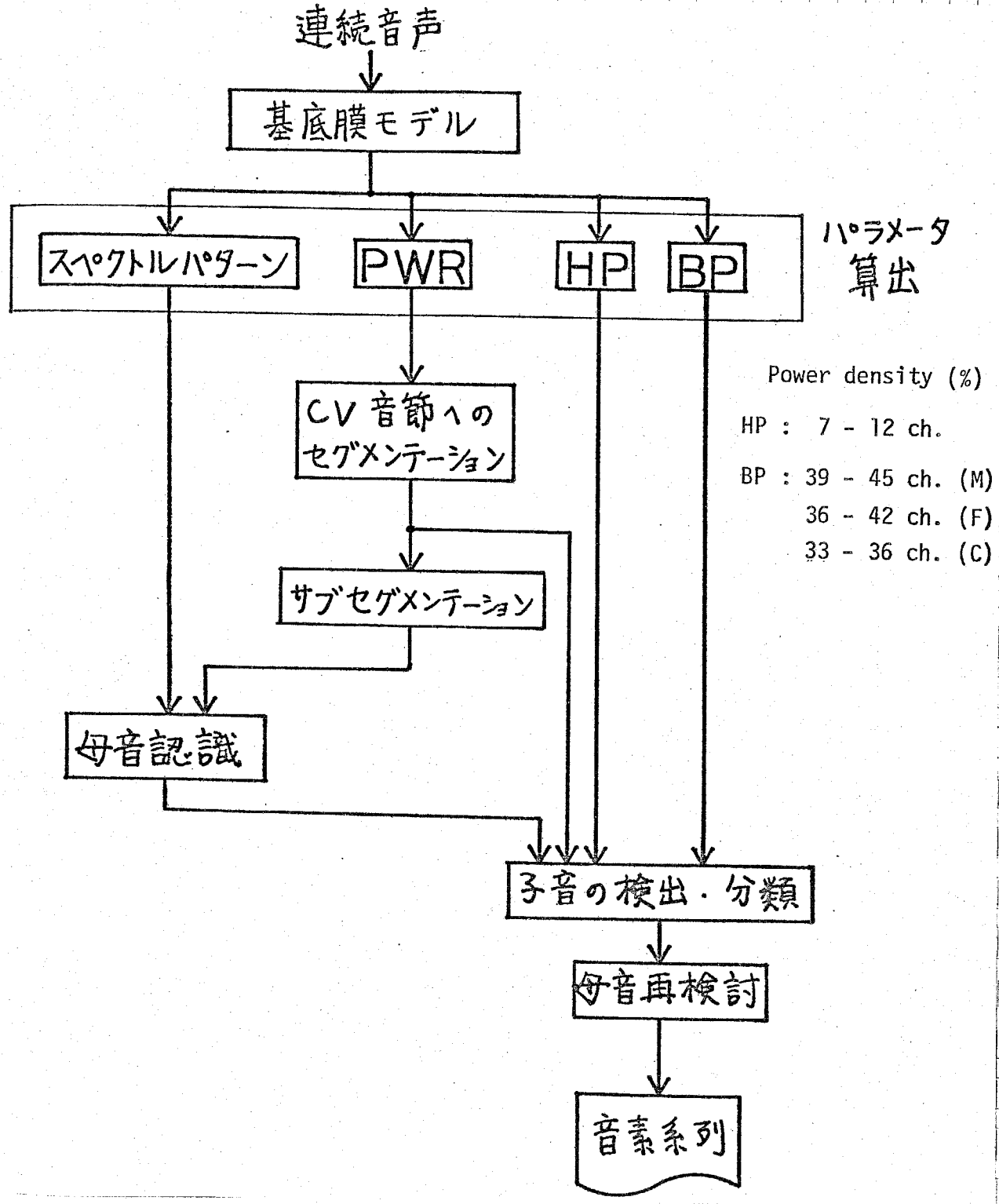


Fig. 5-1 連続音声認識システムの構成．

5.2 セグメンテーション

5.2.1 子音部検出用セグメンテーション

子音部分の検出のために，パワー（PWR）の変化を用いて，連続音声を無音部や，CV，CV*，CVυV，CVN，V，V*（Cは子音，Vは母音，V*は母音連鎖，υは半母音，Nは撥音）等の区間に区切ることを目指す。この操作を「セグメンテーション」と呼び，この操作により区切られた区間を「セグメント」と呼ぶことにする。各フレームごとのパワー（PWR）は，（5－1）式のように定義する。

kチャネルの基底膜出力を $f_k(n)$ とすると，

$$PWR(l) = \frac{1}{N} \sum_{k=1}^{54} \sum_{n=0}^{N-1} f_k(n + (l-1)N)^2 \qquad (5-1)$$

l：フレーム番号　　$N = 512$　　フレーム長 $NT = 12.8$ msec

各セグメント内での母音連鎖中の母音や半母音・撥音の境界の検出は後の処理に任せる。

パワーの変化を用いてセグメンテーションを実行する方法としては，パワーの極小点を検出する方法がよく用いられるが，連続音声中では次の音節への物行がパワーの極小点に達する前に完了している場合も見られ，後に子音の認識等を実行する際に子音の特徴の欠落を招く恐れがある。そこで，パワーの減少している部分に着目し，前後のフレームのパワー（PWR）の比を用いてセグメンテーションを実行する。パワーが減少している部分の多くは，音節間のわたりの部分であり，この部分でセグメンテーションを実行しても音韻認識にはそれほど影響はない。

区切られた各セグメントには，分類のためにセグメント番号を割り当てる。各フレームに対し，セグメント番号を示す変数 SEG#（l）（l：フレーム番号）を対応させ，同一セグメント内のフレームのSEG#（l）は同じ値を持

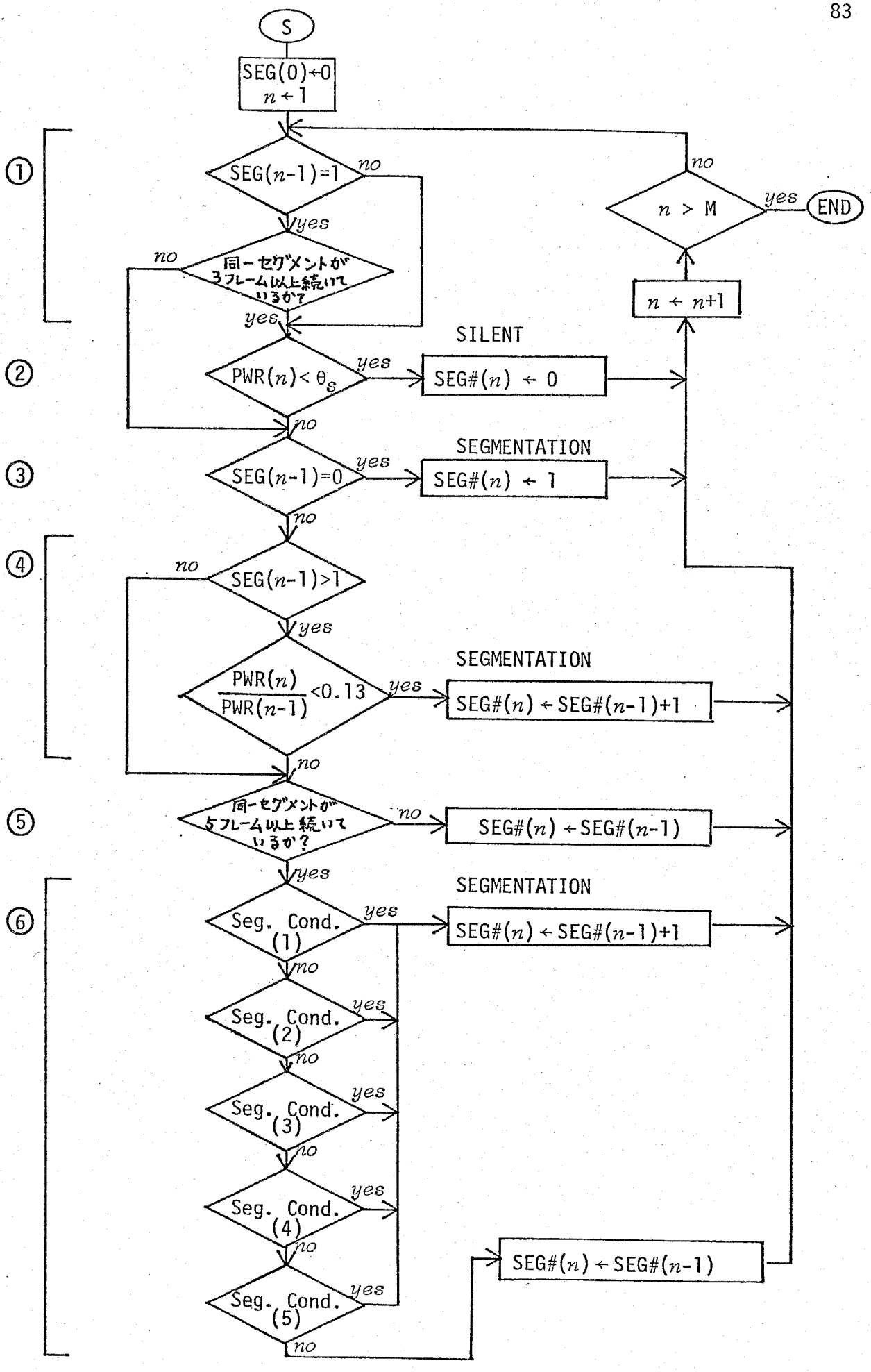


Fig. 5-2 セグメンテーションのアルゴリズム.

たせる。セグメント番号は，無音部（silent）に 0 を割り当て，順次1ずつ増加させ，次の無音部が出現したところで，また0にリセットする。

Fig. 5－2に，セグメンテーションのアルゴリズムのフローチャートを示し，同図にそって各手順を説明する。無音部を決定する閾値 θ_S は，音声信号の最大振幅をA/Dコンバータのフルスケールに合わせて入力した場合 17,000. と固定した。また，他のパワー減少率の閾値は，男性話者3名の発声した各種VCV音節を使い，パワーの変化の観察や切り出し部の試聴実験をくり返して決定した。以下，Fig. 5－2中の手順①～⑥までの手続きを説明する。

①. 無声破裂音等の場合，一度無音部から立ち上がり，次にまたすぐに無音部が出現する場合がある。Fig. 5－3に一例を示すが，4フレーム目に無音部が出現するが，ここでセグメンテーションを実行すると破裂部が欠落してしまい，後に子音認識をする場合に困る。手順①は，それを避けるための操作である。

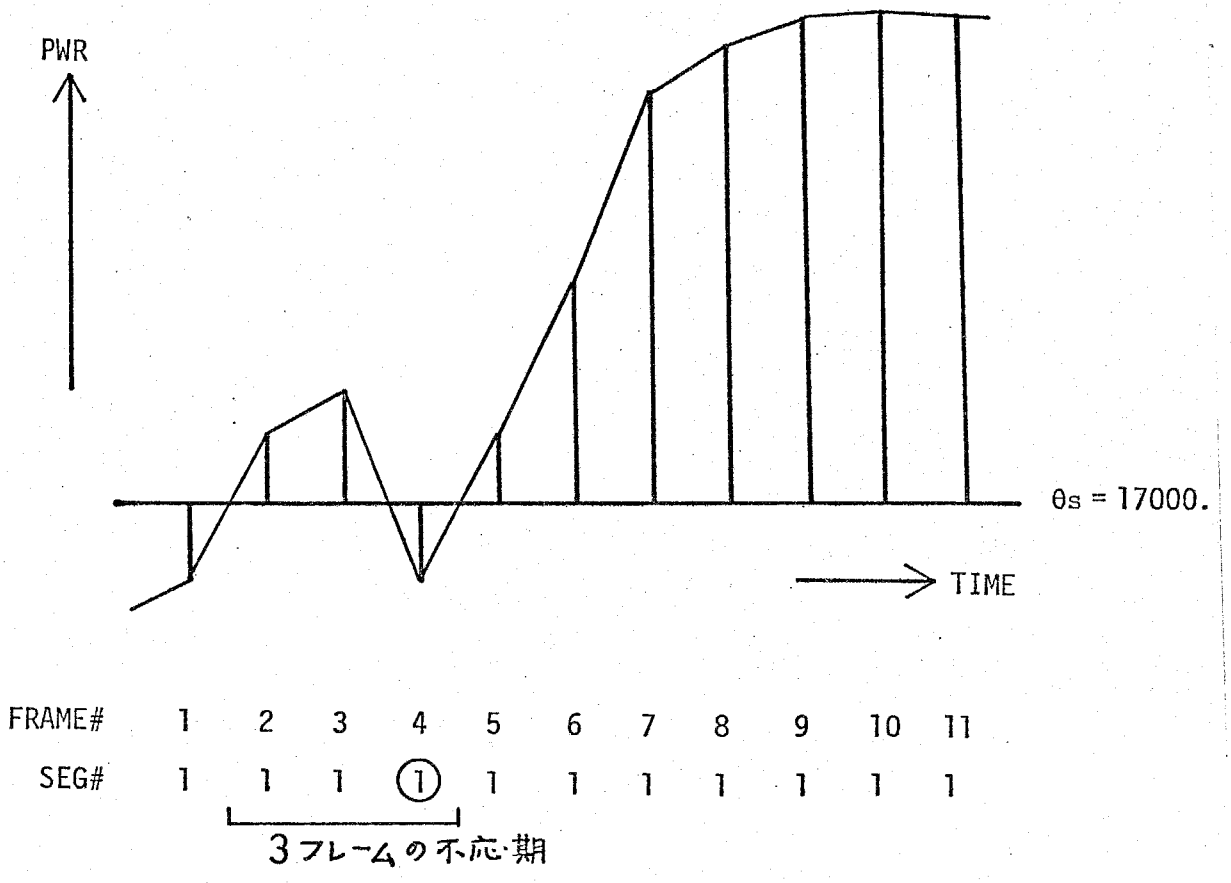


Fig. 5-3 手順①による破裂部欠落の防止.

②. 無音部（silent）検出。

③. 無音部から有音部へ移る点でセグメンテーションをする。

④. 口を閉じたり，声帯振動を停止する時の急激なパワーの変化を検出し，セグメンテーションをする。

⑤. 1つのセグメントは，5フレーム以上（64 msec 以上）持続するとし，それ以下ではセグメンテーションは実行しない。

⑥. パワーの減少傾向が，次の5条件のいずれかを満たした場合，セグメンテーションを実行する。

(1). $\frac{PWR(n)}{PWR(n-1)} < 0.35$

(2). $\frac{PWR(n-1)}{PWR(n-2)} < 1.17$ かつ. $\frac{PWR(n)}{PWR(n-1)} < 0.49$

(3). $\frac{PWR(n-1)}{PWR(n-2)} < 1.17$ かつ, $\frac{PWR(n)}{PWR(n-1)} < 0.56$

かつ, $\frac{PWR(n+1)}{PWR(n)} < 1.16$

(4) $\frac{PWR(n-1)}{PWR(n-2)} < 1.17$ かつ, $\frac{PWR(n)}{PWR(n-1)} < 1.1$

かつ, $\frac{PWR(n)}{PWR(n-3)} < 0.41$ かつ, $\frac{PWR(n+1)}{PWR(n)} < 1.16$

(5) $\frac{PWR(n-2)}{PWR(n-3)} < 1.01$ かつ, $\frac{PWR(n-1)}{PWR(n-2)} < 0.902$

かつ, $\frac{PWR(n)}{PWR(n-1)} < 0.784$ かつ, $\frac{PWR(n+1)}{PWR(n)} < 1.16$

Fig. 5-4にこの5条件でセグメンテーションされる限界のパワー変化の様子を示す。

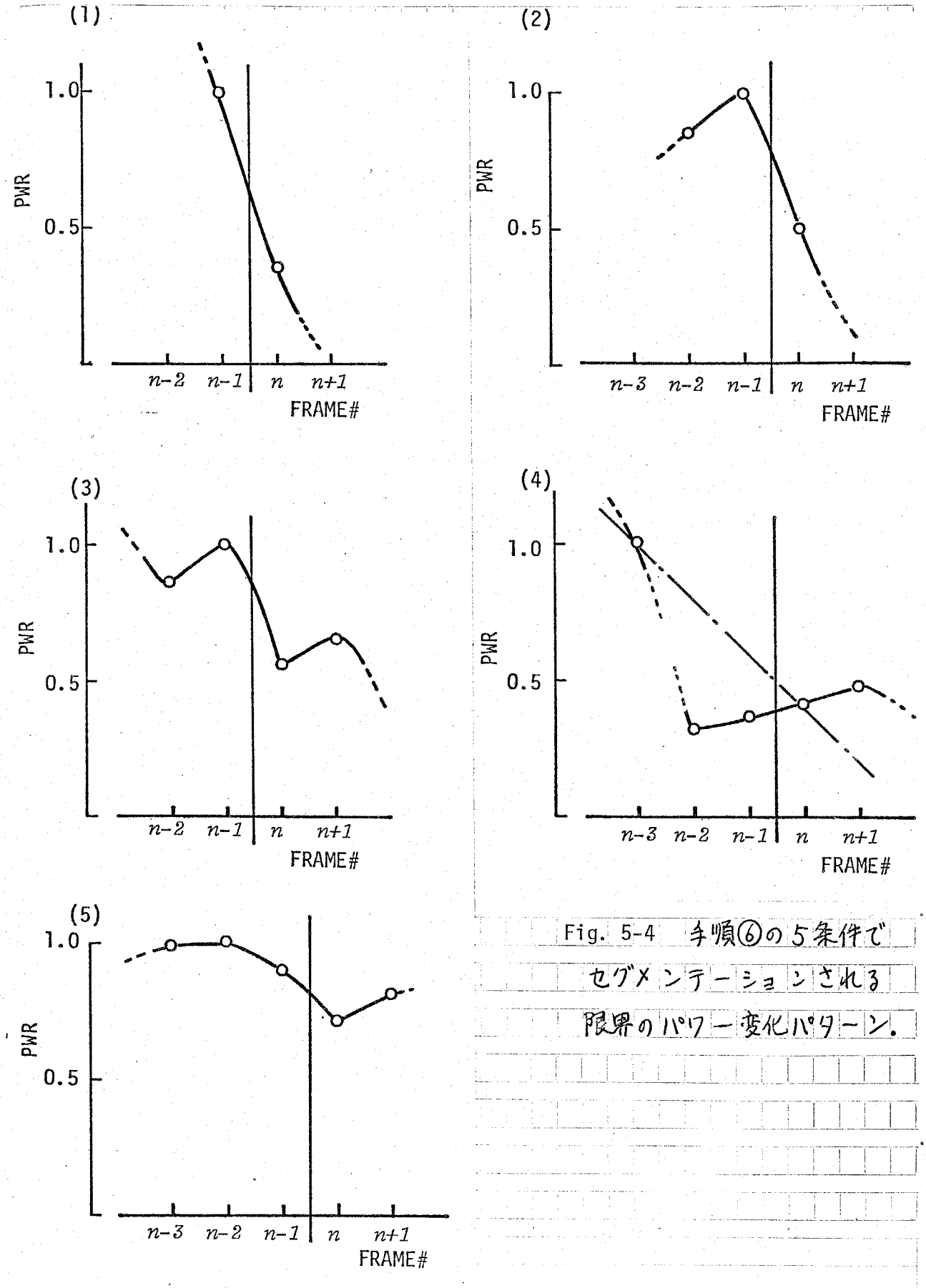


Fig. 5-4 手順⑥の5条件で
セグメンテーションされる
限界のパワー変化パターン。

5.2.2 サブセグメンテーション

5.2.1の手順で求めたセグメント内には，母音連鎖や半母音等が含まれている場合がある。そこで，母音と半母音の境界，母音と母音の境界を区切るために，セグメント内をさらに分割する。この操作を「サブセグメンテーション」と呼び，この操作により区切られた区間を「サブセグメント」と呼ぶことにする。サブセグメンテーションは，以下に述べるようなセグメント内のパワーの極小点を検出することにより行ない，1つの極小点から次の極小点の1つ前のフレームまでを1つのサブセグメントとする。極小点の定義を以下に述べる。

①. セグメントの始点においては，$PWR(n) < PWR(n+1)$ であれば極小点とする。

②. 過去同一サブセグメントが4フレーム以上続いていれば，$PWR(n-1) > PWR(n)$，かつ，$PWR(n) < PWR(n+1)$ ならば，そのフレーム（フレーム番号 n）を極小点とする。

③. 過去同一サブセグメントが3フレーム続き，$PWR(n-1) > PWR(n)$ かつ，$1.15 \leqq \frac{PWR(n+1)}{PWR(n)}$ ならば，そのフレーム（フレーム番号 n）を極小点とする。

④. 過去同一サブセグメントが3フレーム続き，$\frac{PWR(n)}{PWR(n-1)} \leqq 0.979$ かつ $PWR(n) < PWR(n+1) < PWR(n+2)$ ならば，そのフレーム（フレーム番号 n）を極小点とする。

但し，③と④の手続きにおいて，セグメント内の最初の極小点を検出する際には，サブセグメント長に関する制限は無視する。

セグメンテーションの場合と同様に，各サブセグメントに対して，サブセグメント番号を与える。このサブセグメント番号は，1つのセグメントの中だけで有効な値である。各フレームに，サブセグメント番号を示す変数としてSUBSEG#(l)を対応させ，サブセグメントの区別に利用する。セグメントの始点から最初の極小点の直前のフレームまではSUBSEG#=0とし，以下極小点ごとにサブセグメント番号を1ずつ増加していく。極小点がセグメントの最初のフレームにある場合はSUBSEG#=1からはじめる。つまりSUBSEG#=0の区間は，パワーが減少している区間で，直前の音節からのわたりの区間と見なせる。

Fig. 5-5にサブセグメンテーションの手順のフローチャートを示し，Fig. 5-6に，実際に連続音声をセグメンテーション及びサブセグメンテーションした例を示す。

なお，セグメント内の極小点の数によりセグメントを分類し，極小点が1つもないセグメントをType 0のセグメント，極小点が1個のセグメントをType 1のセグメント，2個のセグメントをType 2のセグメント…と呼ぶことにする。また，サブセグメントは，そのサブセグメント番号により，サブセグメント0，サブセグメント1，…と呼ぶことにする。

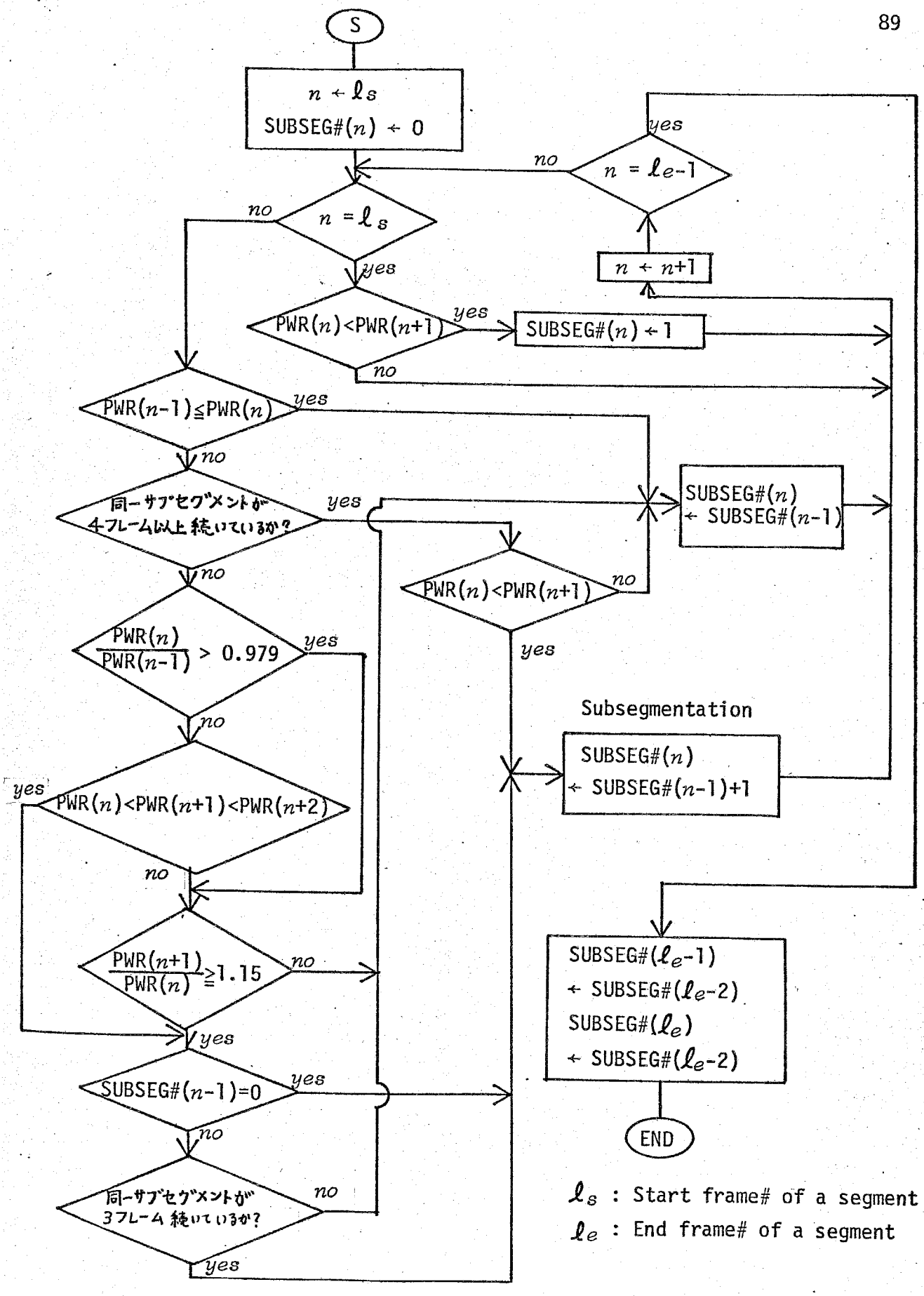


Fig. 5-5 サブセグメンテーションのアルゴリズム.

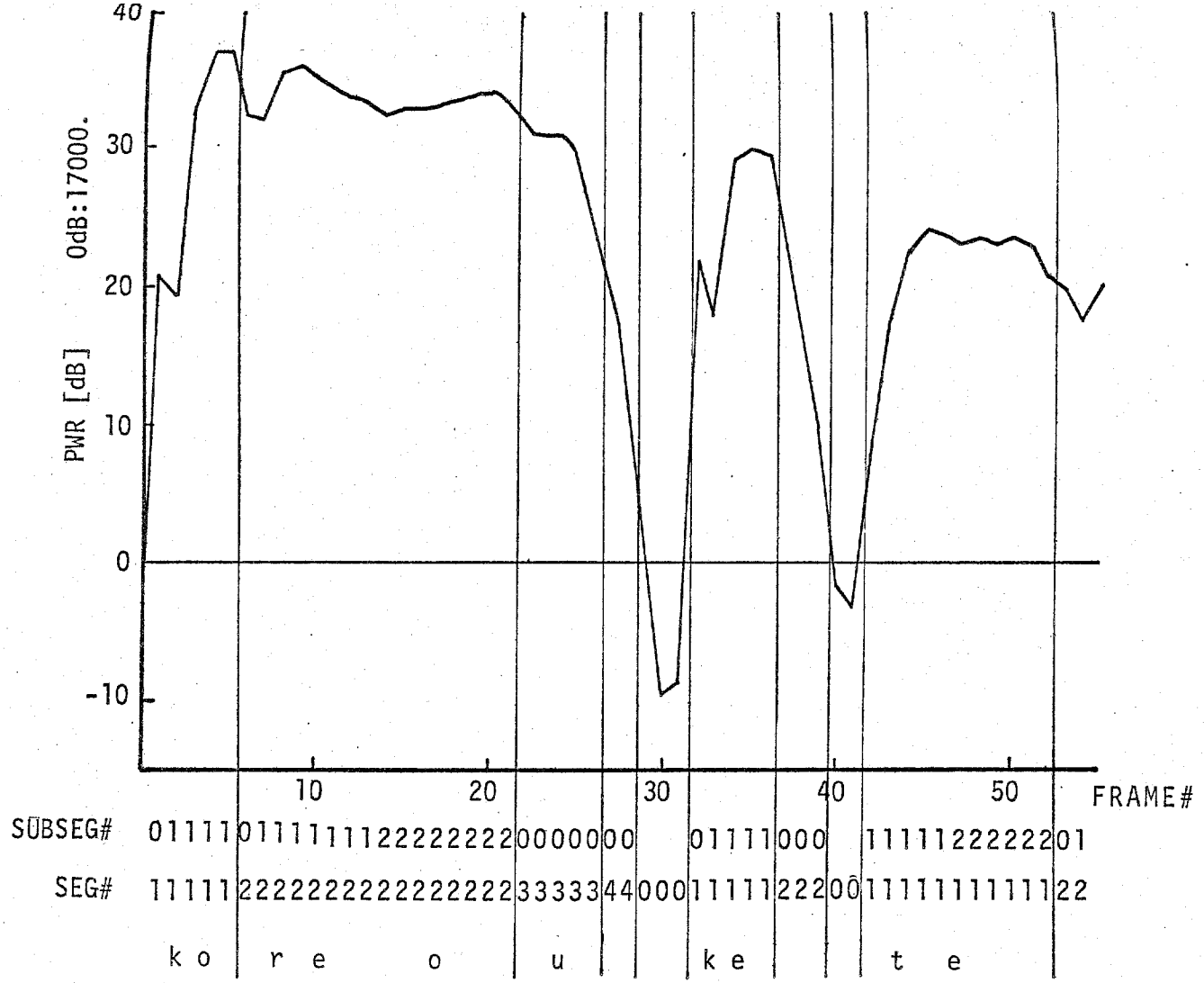


Fig. 5-6 セグメンテーション及びサブセグメンテーションの一例.

5．3　母音認識と子音の分類

5．2節で示したアルゴリズムにより求められたセグメント，サブセグメントをもとにして，母音の認識と子音の検出・分類を行なう。以下の処理は1つのセグメントを単位として実行する。無音区間のセグメントは当然のことながら5，3フレーム以下の持続時間のセグメントも処理対象からはずす。3フレーム以下のセグメントは，Fig．5－2の手順④からわかるように，口を閉じたり声帯振動を停止したりする場合に現われるので音韻情報は含んでいないからである。

5．3．1　母音認識

セグメント内のすべてのフレームに対して，そのスペクトルパターンを用い4章で提示した母音認識を実行し，認識結果として得られた母音カテゴリの系列（母音でないと判断された場合はシンボル「？」を割り当てる），母音認識の第1段階で距離が最小となったカテゴリの系列，及び第2段階へ移行する候補カテゴリ数の系列の3系列を得る。以下，これらの系列をそれぞれ「母音カテゴリ系列」，「距離最小系列」，「母音候補カテゴリ数系列」と呼ぶ。母音候補カテゴリ数系列は「子音の検出・分類」に利用するもので，母音決定には直接関係はない。

母音カテゴリ系列を得たのちに，各サブセグメント内において音韻クラス間の多数決をとり，そのサブセグメントの母音を決定する。ここで多数決をとる音韻クラスは，5母音と非母音クラス／？／の6クラスとする。以下，多数決のとり方を説明する。

（1）．Type 0 セグメントの場合

極小点が存在しないので，全フレームにわたり多数決をとる。その結果，

①．最高得票クラスが1つであれば，その音韻クラスのシンボル（A，I，U，E，O，？）を各フレームに割り当てる。

②. 同票のクラスが2つの場合，そのうちの1クラスが非母音クラスであれば，各フレームにシンボル「?」を割り当てる。2つのクラスが両方とも母音クラスであれば，距離最小系列を使い多数決をとり，その結果得票の多い母音クラスの方を採用する。

③. 同票のクラスが3つ以上の場合，非母音であると判定し，各フレームにシンボル「?」を割り当てる。

（2）. Type 0 以外のセグメントの場合.

各サブセグメント単位で以下の処理を実行する。

A. サブセグメント0（SUBSEG# = 0の区間）の場合.

サブセグメント長が6フレーム以上の場合，Type 0 セグメントと同様の処理を実行する。5フレーム以下の場合は，わたりの部分と見なし認識対象からはずす。

B. サブセグメント1（SUBSEG# = 1の区間）の場合.

サブセグメント長が3フレームの場合は，最初の1フレームを除き，サブセグメント長が4フレーム以上の場合は最初の2フレームを除き，音韻クラス間の多数決をとり，そのサブセグメントの全フレームにその音韻クラスのシンボルを割り当てる。最初の2フレームを除くのは，セグメントの定義によりサブセグメント1の場合は先頭が子音の確率が高いからである。多数決の結果，同票のクラスが2つ以上存在した場合は，除外したフレームのうち直前の1フレームを加え多数決をとる。それでもなお決定できない場合は，非母音と決定し全フレームにシンボル「?」を割り当てる。

C. 他のサブセグメント（SUBSEG# ≧ 2 の区間）の場合.

各サブセグメントの最初の1フレームを除き，音韻クラス間の多数決をとる。同票のクラスがあれば最初の1フレームを加え多数決をとりなおす。それでもなお決定できない場合は非母音と見なし，シンボル「?」を全フレームに割り当てる。

Fig. 5-7に実例を示す。

なお，本章で使用した母音認識システムは，4章で提示した単母音認識システムの第2段階の検証アルゴリズムに多少の変更を加えたものであるが，探索木の枝の数はむしろ減少している。Fig. 5-8 ～ Fig. 5-22にこの連続音声用母音認識システムの各カテゴリの検証アルゴリズムを示すが，単母音認識システムからの変更箇所は図中（＊）で示されている。また，第1段階で用いる参照パターンは単母音認識に用いたものをそのまま使用した。

FRAME#	SEG#	SUBSEG#	NN	Lmin	Vowel	RESULT	
1	1	0	5	CO	CO	( )	
2	1	1	8	MO	MO	(O)	k
3	1	1	7	ME	MO	(O)	
4	1	1	8	MO	MO	(O)	o
5	1	1	8	ME	MO	(O)	
6	2	0	6	ME	MU	( )	
7	2	1	7	ME	MO	(E)	
8	2	1	5	ME	ME	(E)	r
9	2	1	5	ME	MO	(E)	
10	2	1	5	ME	ME	(E)	e
11	2	1	5	ME	ME	(E)	
12	2	1	5	ME	ME	(E)	
13	2	1	5	ME	ME	(E)	
14	2	2	7	ME	MO	(O)	
15	2	2	7	ME	MO	(O)	
16	2	2	8	MO	MO	(O)	
17	2	2	8	MO	MO	(O)	
18	2	2	8	MO	MO	(O)	o
19	2	2	8	MO	MO	(O)	
20	2	2	7	MO	MO	(O)	
21	2	2	8	MO	MO	(O)	
22	3	0	6	MO	MO	(U)	
23	3	0	6	MO	MO	(U)	
24	3	0	6	MU	MU	(U)	u
25	3	0	5	MU	MU	(U)	
26	3	0	3	MU	MU	(U)	
27	4	0					
28	4	0					
29	0				SILENT		
30	0				SILENT		
31	0				SILENT		
32	1	0	0	MA	?	( )	
33	1	1	0	FA	?	(E)	k
34	1	1	2	ME	ME	(E)	
35	1	1	4	ME	ME	(E)	e
36	1	1	5	ME	ME	(E)	
37	2	0					
38	2	0					
39	2	0					
40	0				SILENT		
41	0				SILENT		
42	1	1	0	CU	?	(E)	
43	1	1	4	ME	ME	(E)	
44	1	1	3	ME	MU	(E)	t
45	1	1	6	ME	ME	(E)	
46	1	1	4	ME	ME	(E)	e
47	1	2	4	ME	ME	(E)	
48	1	2	4	ME	ME	(E)	
49	1	2	5	ME	ME	(E)	

NN : 母音候補カテゴリ数系列
Lmin : 距離最小系列
Vowel: 母音カテゴリ系列

Fig. 5-7 母音認識の実例.

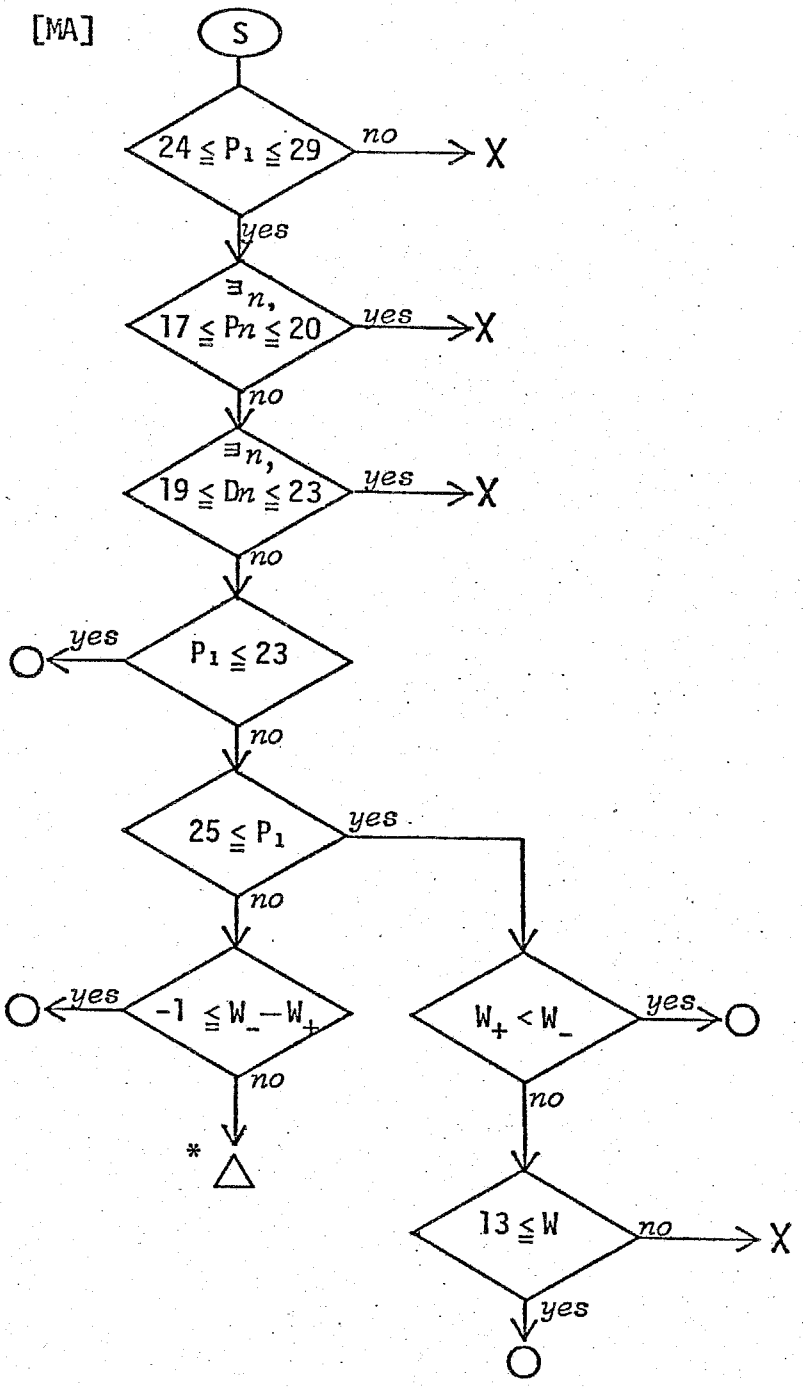


Fig. 5-8　連続音声用[MA]の検証アルゴリズム.

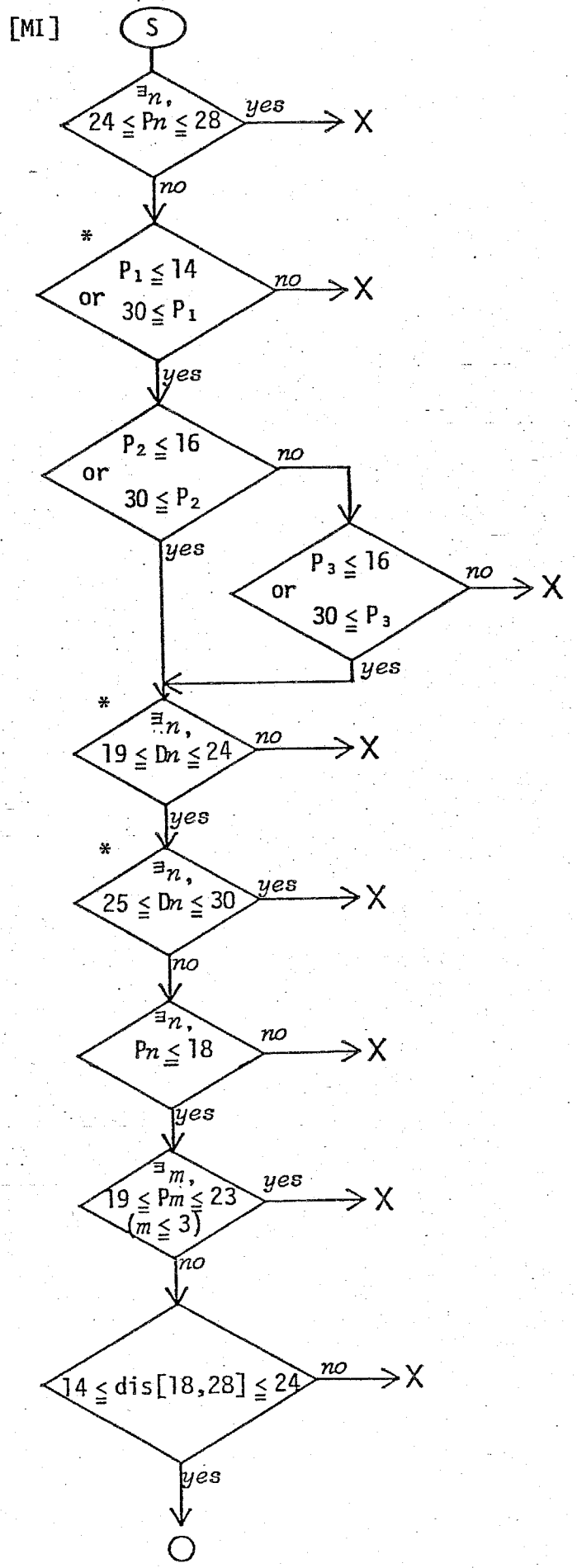


Fig. 5-9　連続音声用[MI]の検証アルゴリズム.

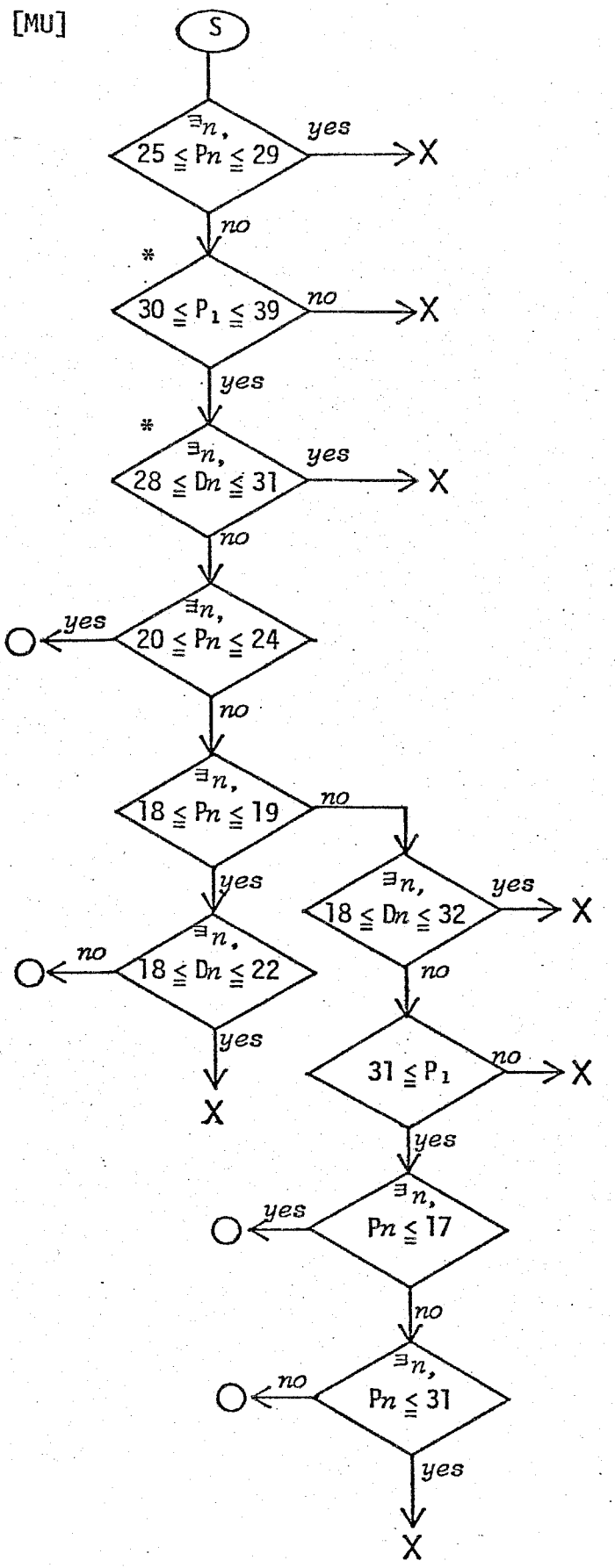


Fig. 5-10 連続音声用[MU]の検証アルゴリズム.

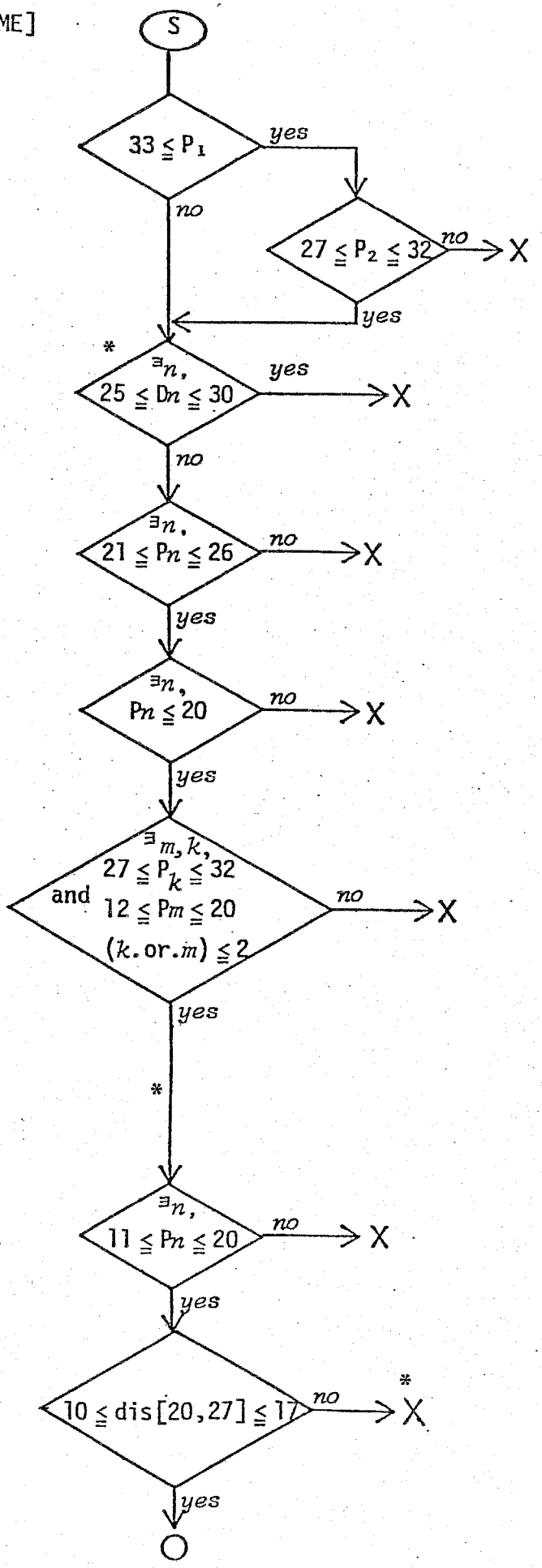


Fig. 5-11 連続音声用[ME]の検証アルゴリズム.

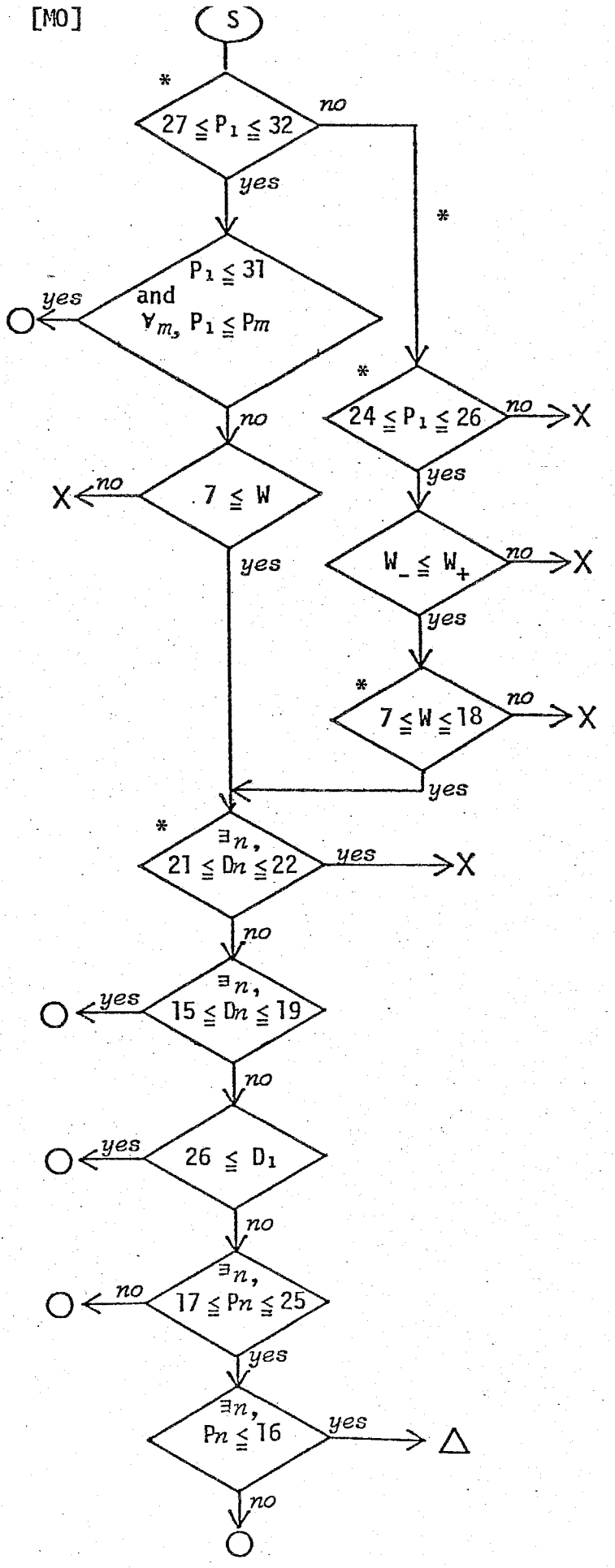


Fig. 5-12 連続音声用[MO]の検証アルゴリズム.

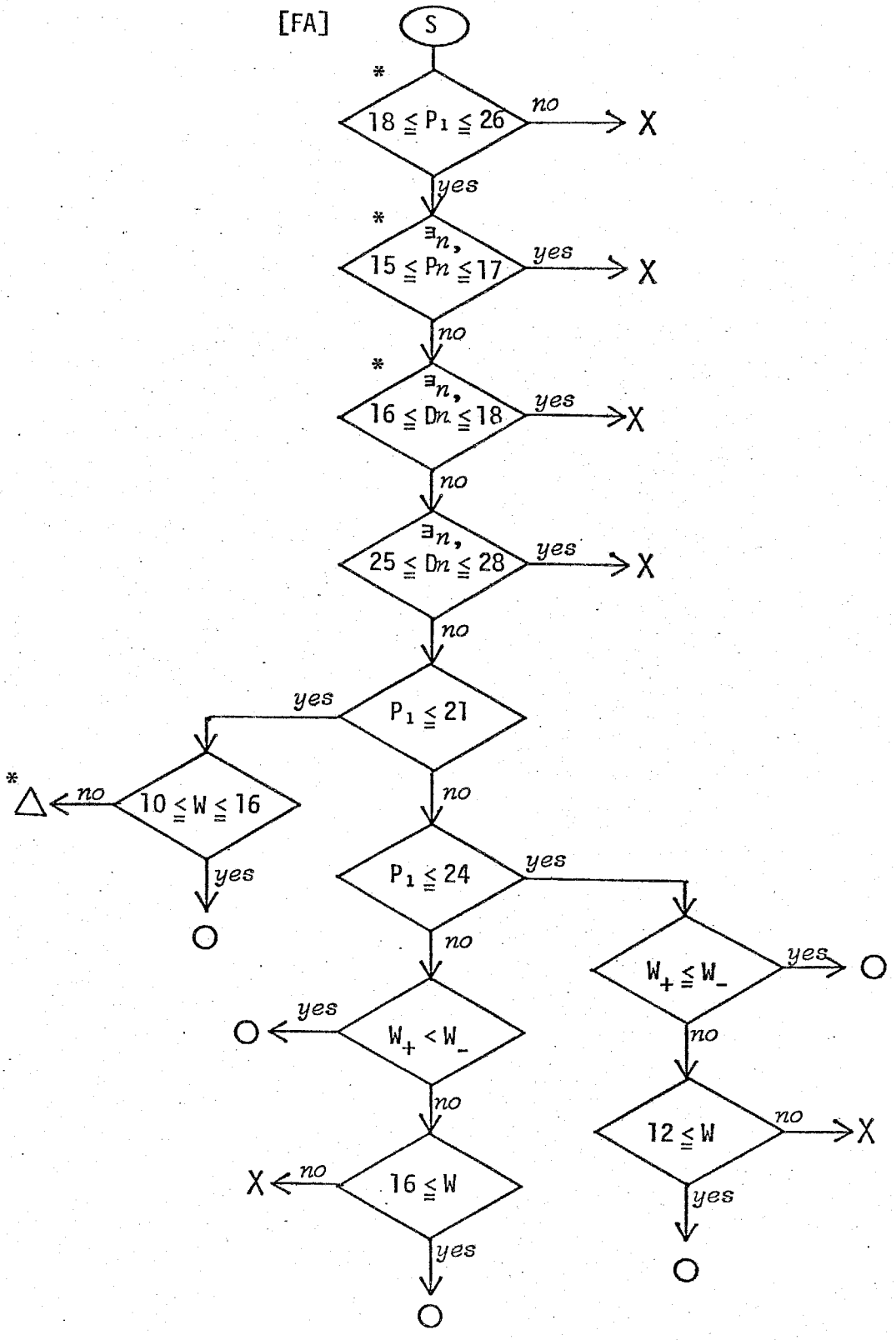


Fig. 5-13 連続音声用[FA]の検証アルゴリズム.

[FI]

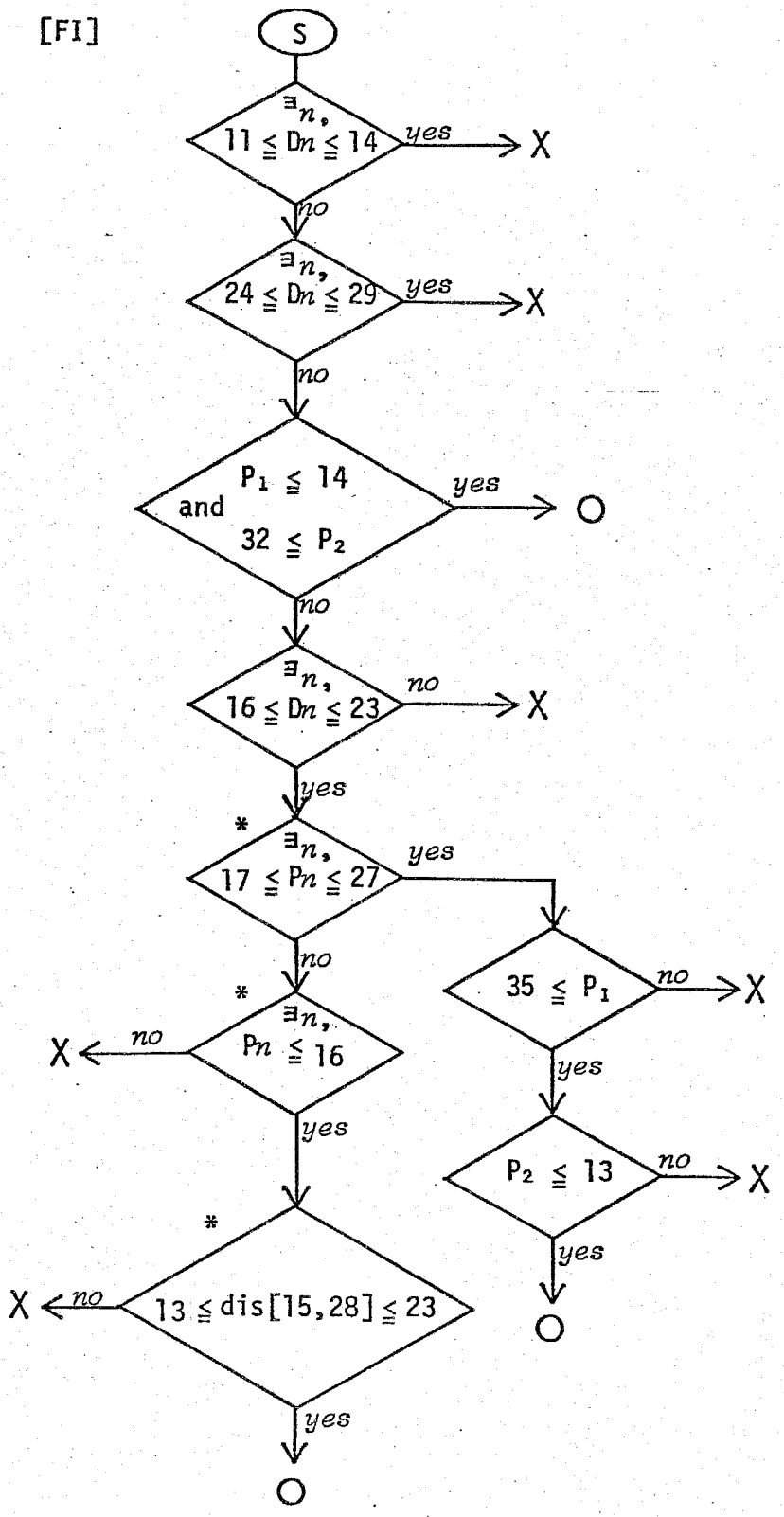


Fig. 5-14　連続音声用[FI]の検証アルゴリズム.

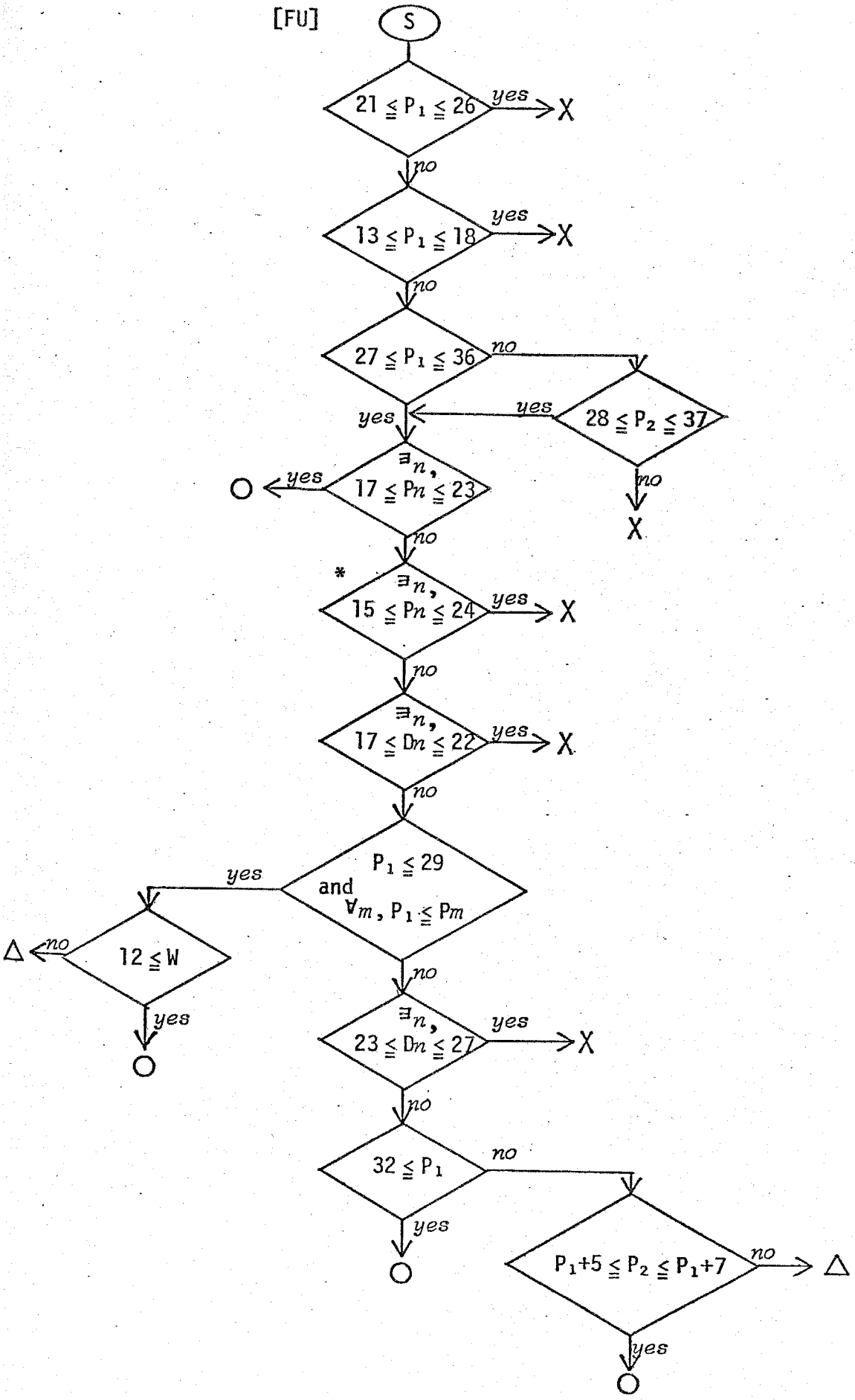


Fig. 5-15　連続音声用 [FU] の検証アルゴリズム.

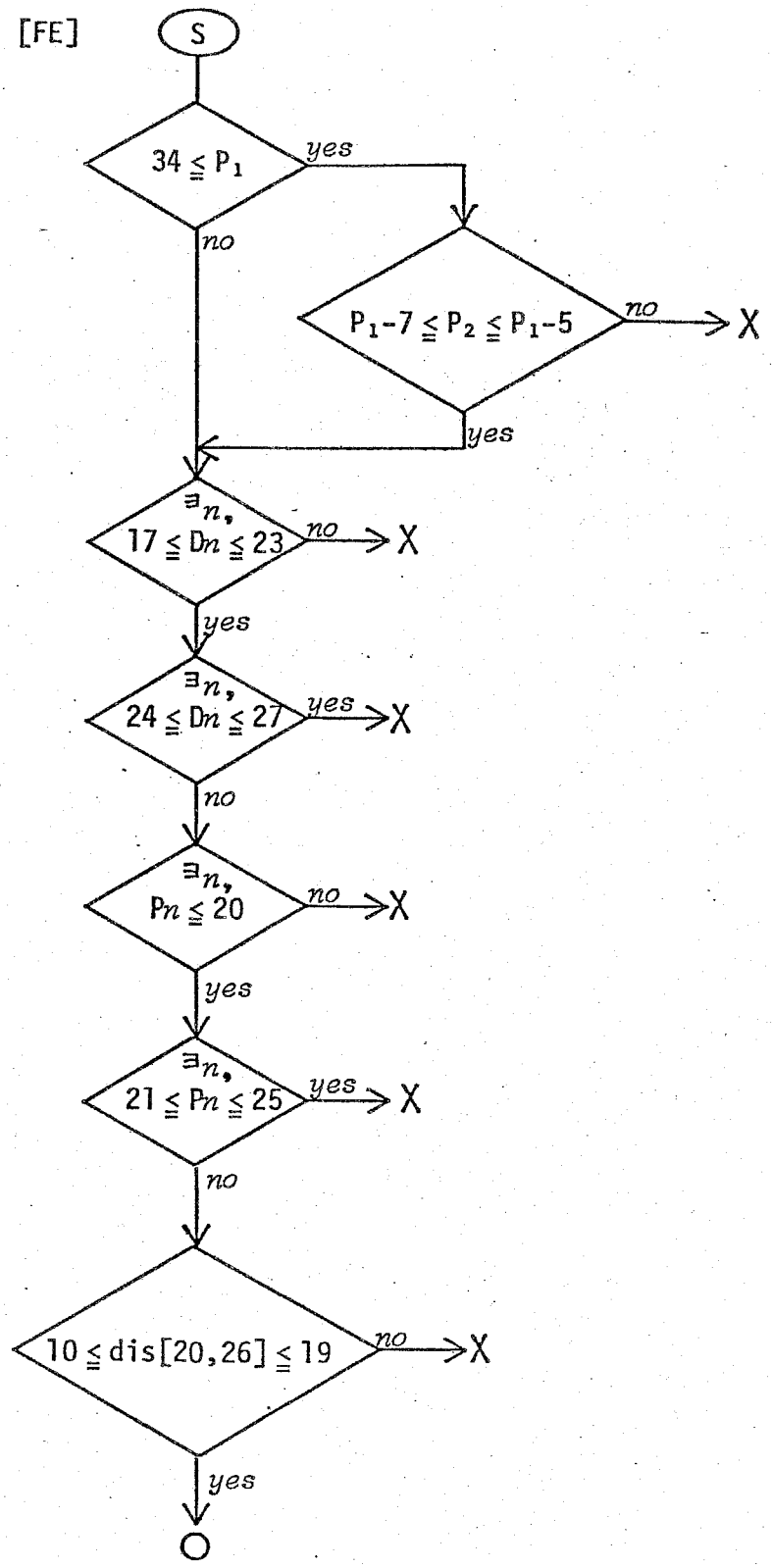


Fig. 5-16　連続音声用 [FE] の検証アルゴリズム.

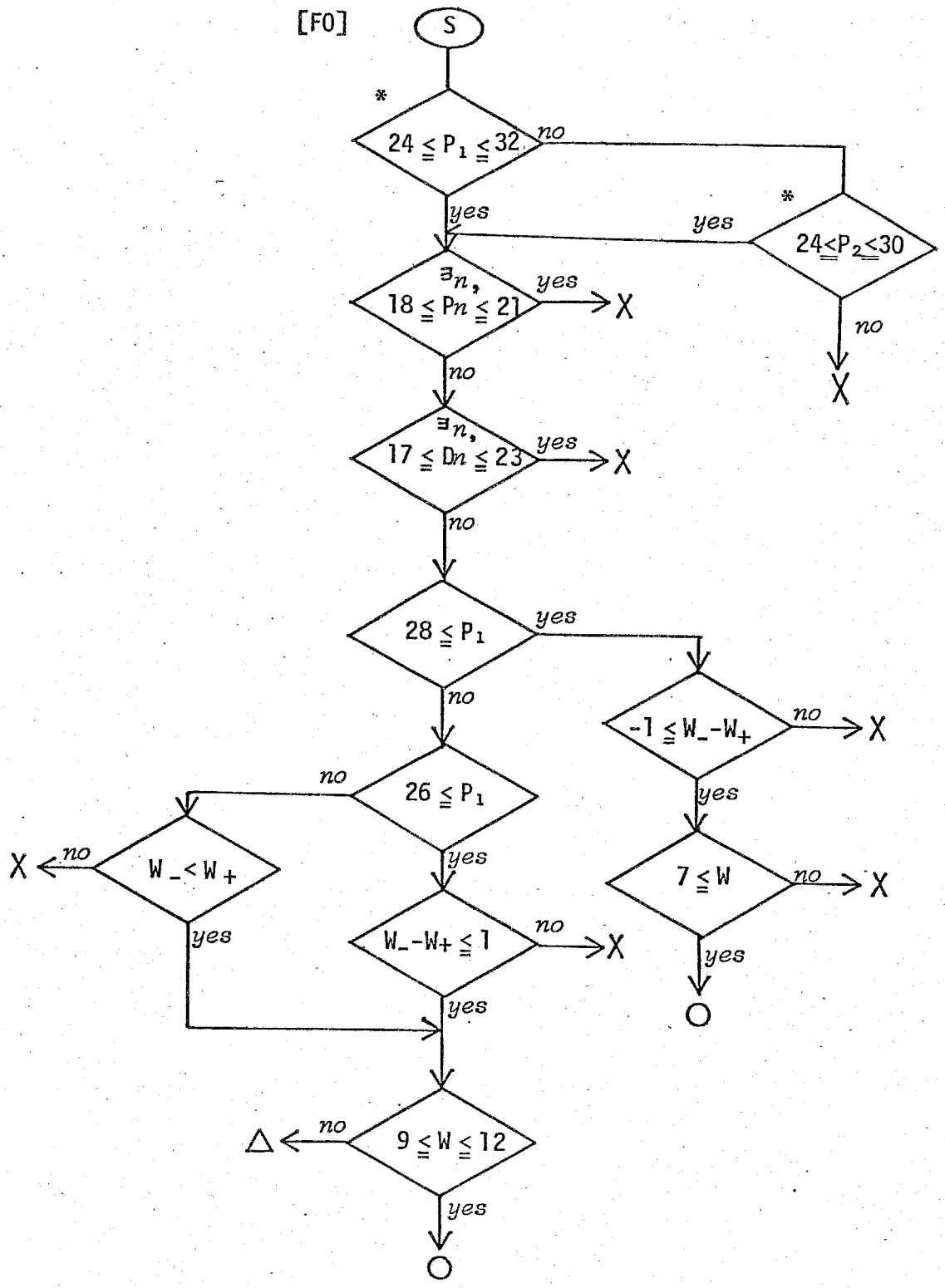


Fig. 5-17　連続音声用 [FO] の検証アルゴリズム.

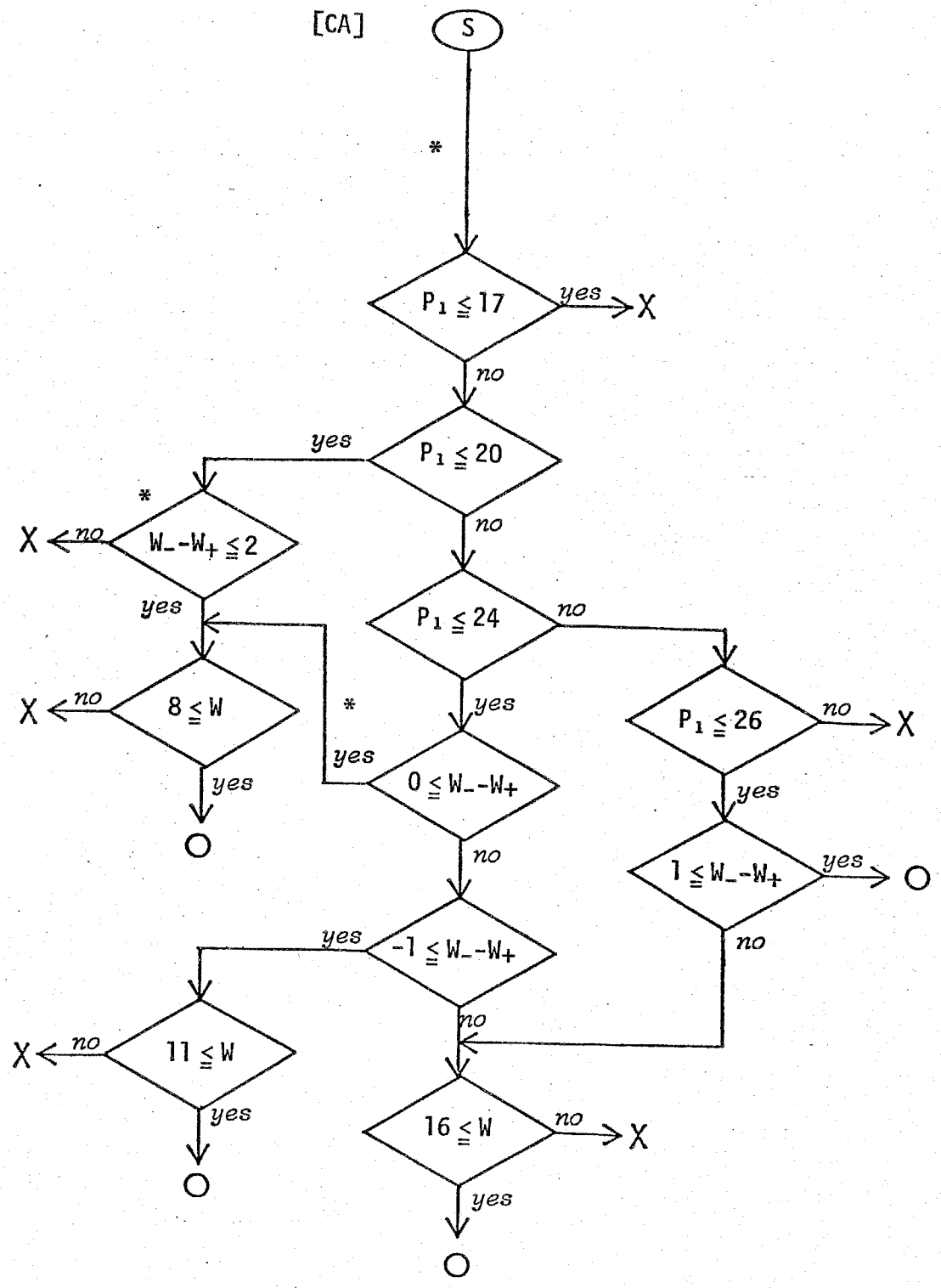


Fig. 5-18　連続音声用 [CA] の検証アルゴリズム.

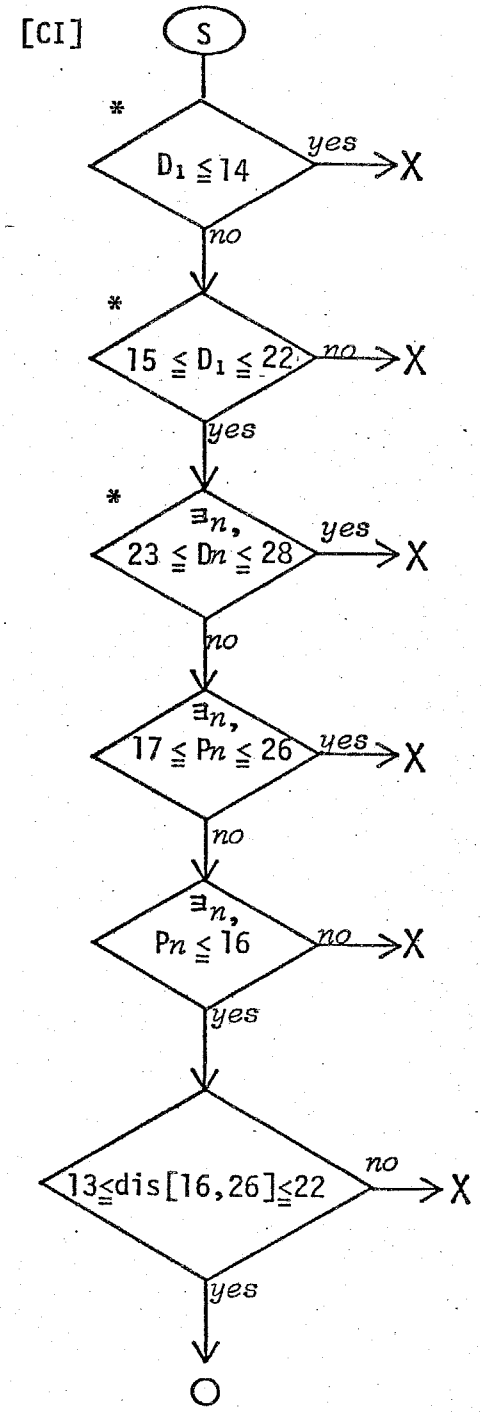


Fig. 5-19　連続音声用 [CI] の検証アルゴリズム.

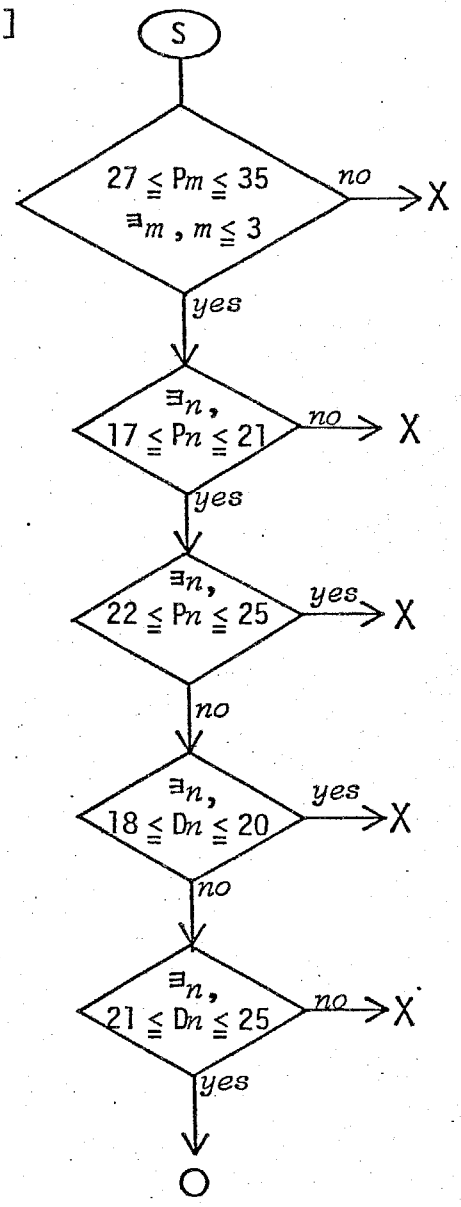


Fig. 5-20　連続音声用 [CU] の検証アルゴリズム.

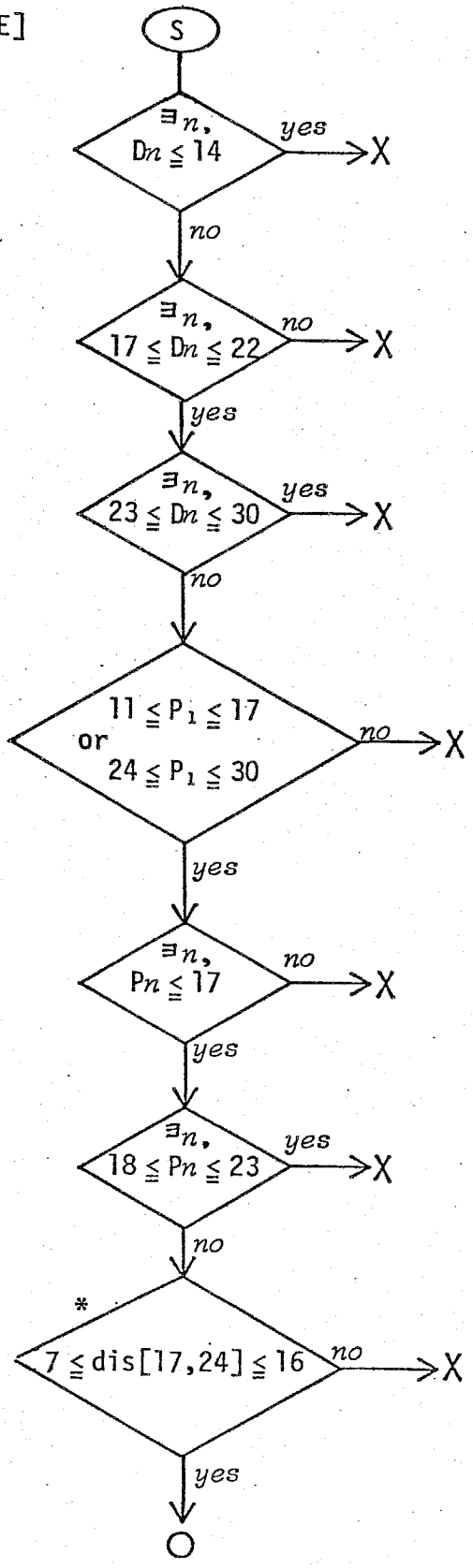


Fig. 5-21　連続音声用[CE]の検証アルゴリズム.

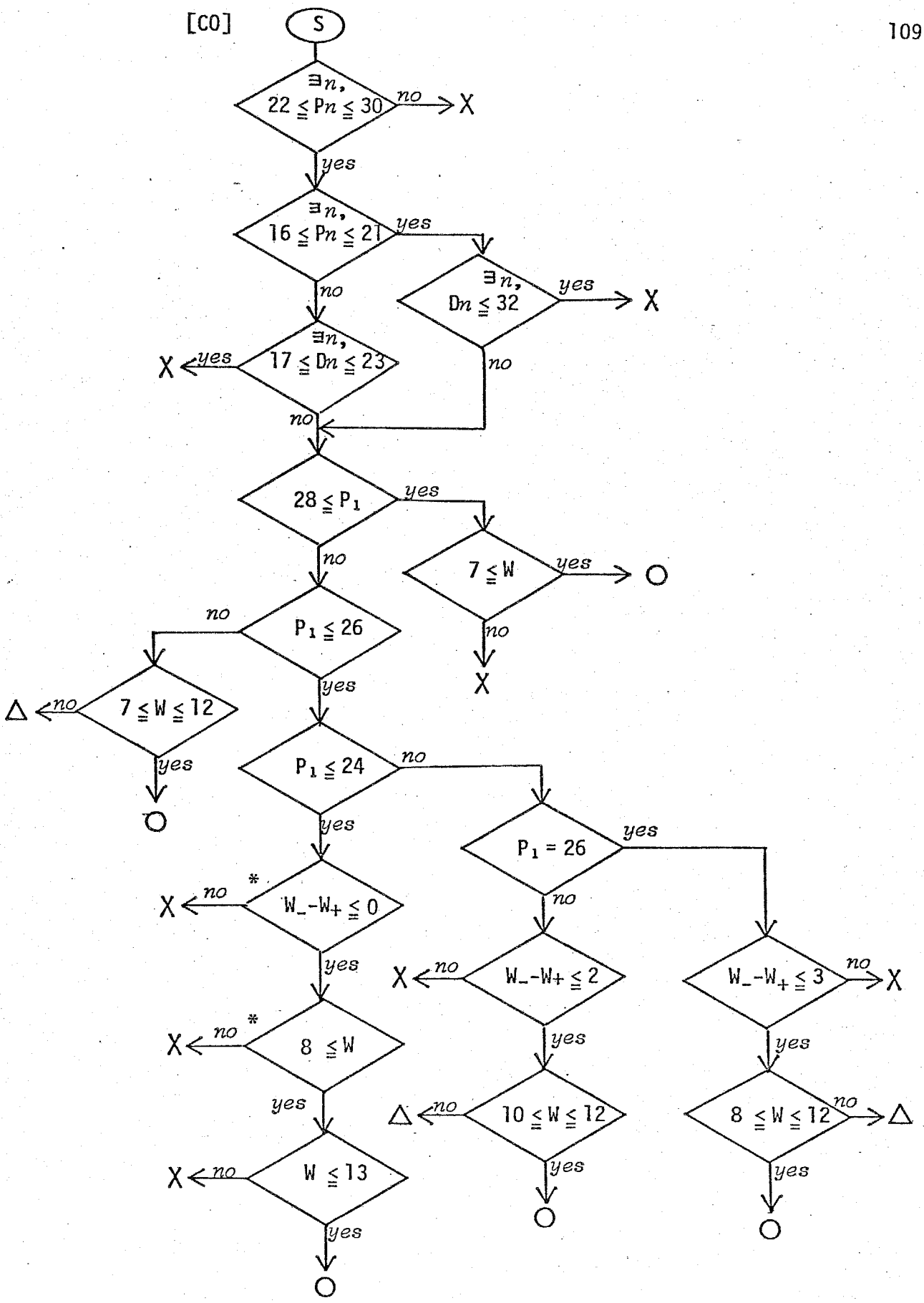


Fig. 5-22　連続音声用 [CO] の検証アルゴリズム.

5.3.2 子音の検出・分類

子音の検出・分類には，母音認識の結果と各フレームのパワー（PWR）の他に，次の2つのパラメータを使用する。1つは摩擦音の検出に用いるためのパラメータで，7チャネル（CF 7.13 kHz）から12チャネル（CF 4 kHz）の6チャネルにわたってのパワーの集中度を百分率で表わしたもので，変数 HP(n)（nは各セグメントの先頭から数えたフレーム数）で表記する。もう1つのパラメータは，有声音と無声音の区別をするためのもので，ピッチ周波数周辺の7チャネルにわたるパワーの集中度を百分率で表わしたもので 変数 BP(n)（nは各セグメントの先頭から数えたフレーム数）で表記する。ピッチ周波数がどのチャネルに出現しているかは直接求めず，連続音声データの最初の有音区間20フレームの母音認識結果の母音カテゴリ系列から，属性クラスに関する多数決をとり，「男性」と判断されれば39チャネル（CF 176 Hz）から45チャネル（CF 88 Hz）の7チャネル，「女性」なら36チャネル（CF 250 Hz）から42チャネル（CF 125 Hz），「子供」ならば33チャネル（CF 353 Hz）から39チャネルのパワーの集中度を計算する。HP(n)及びBP(n)は，スペクトルパターン $X(n) = (x_1(n), x_2(n), \cdots x_{54}(n))$ から直接（5-2），（5-3）式で計算できる。

$$HP(n) = \sum_{k=7}^{12} x_k(n)^2 \times 100 \quad (\%) \qquad (5-2)$$

$$BP(n) = \sum_{k=g}^{g+6} x_k(n)^2 \times 100 \quad (\%) \qquad (5-3)$$

「男性」 $g = 39$

「女性」 $g = 36$

「子供」 $g = 33$

子音は，大まかに 無声破裂音（p，t，k），無声摩擦音（s，sh，ch，ts，h），有声摩擦音（z），その他の有声子音（N，m，n，r，b，d，g，y，w）の4クラスに分類することを目指し，このほかに補足的に，k，h，（ch，ts），撥音 N の検出を試みた。Fig. 5－23 に，子音の検出・分類のフローチャートを示すが，以下同図（1）～（8）の処理手順について詳述する。子音が検出された場合，子音区間の各フレームに対して各子音群のシンボルを割り当てるが，その際，母音認識で割り当てられた母音シンボルがある場合は，子音のシンボルに書き換えるという操作を実行する。各子音群のシンボルは，以下のように決定した。

	シンボル
無声破裂音（p，t，k）	「TK」
無声摩擦音（s，sh，ch，ts，h）	「FR」
有声摩擦音（z）	「Z」
他の有声子音（N，m，n，r，b，d，g，y，w）	「VC」
k	「K」
h	「H」
破擦音（ch，ts）	「CT」
撥音 N	「N」

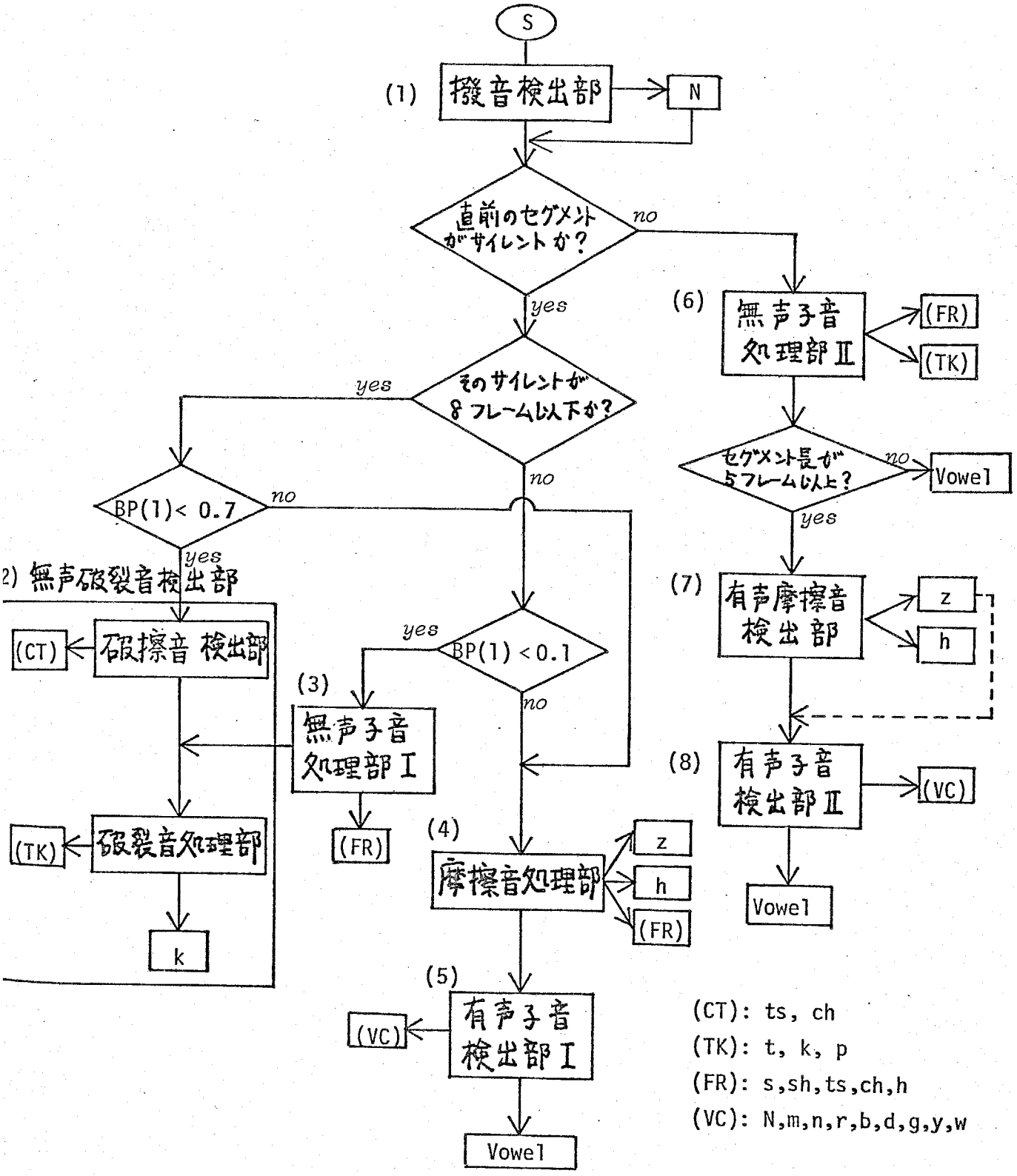


Fig. 5-23 子音の検出・分類のフローチャート.

セグメントの定義から，各セグメントには撥音Nを除き2つ以上の子音は存在しないと仮定して，各セグメントで1つの子音が検出されたならば，子音の検出処理は終了する。セグメントにより検出できない半母音は，母音として処理する。

Fig. 5-29 ～ Fig. 5-35に実際の子音検出例を示すので，以下各項でそれらを参考にしながら子音検出手順を説明する。

（1）. 撥音検出部

撥音Nは，セグメンテーションの結果，Type 0のセグメントやサブセグメント0（以上2つの区間はパワーが減少傾向にある区間），及びType 2以上のセグメントの最終サブセグメントに出現する頻度が高い。そこで以上の3区間において撥音の検出を実行する。撥音のスペクトルパターンは，母音/i/や/u/のスペクトルパターンに近いことから，先の母音認識の結果が/i/又は/u/であるか，或いは非母音/?/であっても距離最小系列の多数決の結果が/i/又は/u/である場合についてのみ，撥音検出を実行する。サブセグメント0の区間は，サブセグメント長が6フレーム以上でなければ，わたりの区間とみなして撥音検出は実行しない。以下，それぞれの場合についての検出条件を述べる。

A. Type 0のセグメント

①. セグメント長が15フレーム以上続けば撥音とみなし，シンボル「N」を全フレームに割り当てる。

②. HP(n)がすべてのフレームにわたり7%以下であり，BP(1)（セグメントの先頭フレームのBP）が15%以上でかつ，BP(Ls)（Lsはセグメント長，BP(Ls)はセグメントの最終フレームのBP）が10%以上であれば，そのセグメントは撥音であると見なす。

B．サブセグメント 0

BP(1) ≧ 15 (%) であり，サブセグメントの最終フレームのBPが10%以上であり，なおかつそのサブセグメント内の全フレームにわたりHP(n)が8%以下であれば，そのサブセグメントを撥音とみなす。

C．Type 2 以上のセグメントの最終サブセグメント（cf. Fig. 5－29）

サブセグメント長が4フレーム以上であり，サブセグメントの先頭のBPが15%以上，かつ BP(Ls) ≧ 10 (%) であり，さらにそのサブセグメント内のすべてのHP(n)が5%以下であれば，そのサブセグメントを撥音とみなす。

（2）．無声破裂音検出部

無声破裂音は，「声帯振動停止 →破裂」という過程で発声されるので，連続音声中でも，直前に短い無音部（silent）を伴って出現する場合が多い。そこで，直前に8フレーム以下の無音部を伴い，かつ，そのセグメントの先頭のBP(1)が0.7%未満の場合について，無声破裂音と見なし，子音の分類を実行する。破擦音と呼ばれる ch (tʃ)，ts に関しても同様に破裂的傾向があるので，この処理部で検出できる。Fig. 5－24に処理アルゴリズムのフローチャートを示し，以下各部分について説明する。

①．破擦音検出部（cf. Fig. 5－29）

無声摩擦音は，無声部が長く続き，かつスペクトルパターンが母音のパターンと大きく異なるために，母音認識部で得られた母音候補カテゴリ系列を見ると，母音候補カテゴリ数（以下，NNと表わす）が0となる。そこで，セグメントの先頭から NN=0 かつ BP(n) < 0.1 (%) であるフレームが4フレーム以上続けば，破擦音（ch，ts）であると判定し，その該当フレームにシンボル「CT」を割り当てる。さもなくば，次

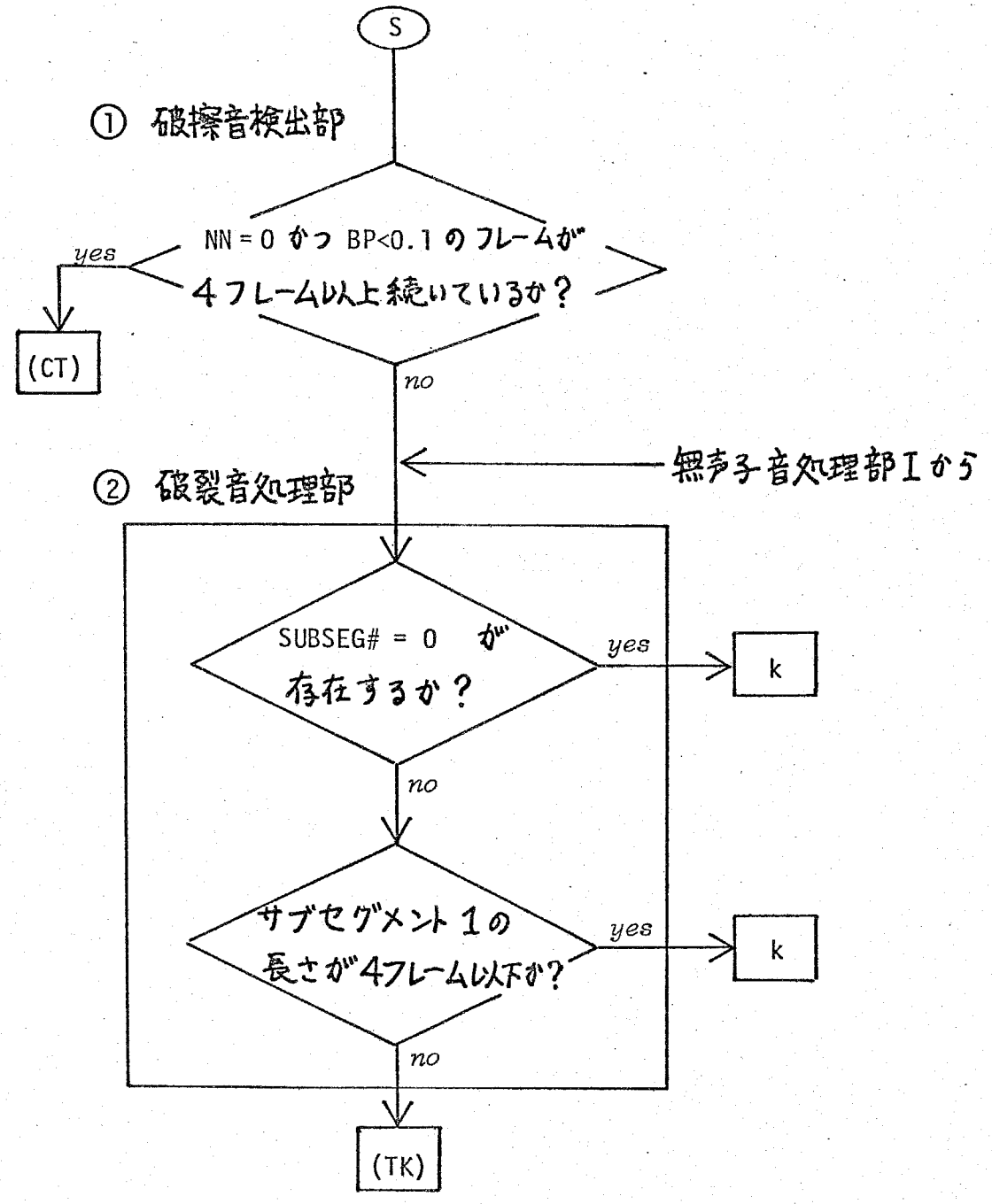


Fig. 5-24 無声破裂音検出部の処理アルゴリズム.

の破裂音処理部へ進む。

②. 破裂音処理部（cf. Fig. 5－29，Fig. 5－34）

無声破裂音 /p/，/t/，/k/ の中で，/k/ だけは破裂直後にパワーの減少が見られ，パワーの極小点が存在することが多い。そこで，サブセグメント0が存在するか，或いは，短いサブセグメント（4フレーム以下）が存在する場合は /k/ であると判断する。この場合，サブセグメント0が存在する場合はサブセグメント0の区間にシンボル「K」を割り当て，サブセグメント1の場合は，先頭の2フレームにシンボル「K」を割り当てる。この2つの場合以外は，無声破裂音のシンボル「TK」をセグメントの先頭の2フレームに割り当てる。

以下の（3）～（5）の各処理部は，直前に無音部を伴う場合，つまり文頭或いは文節の先頭での子音検出を実行する。有声・無声の判断は，セグメントの先頭フレームのBPの値（BP(1)）をもって行ない，BP(1)＜0.1（%）であれば無声子音と判断し，「無声子音処理部I」へ移る。

（3）. 無声子音処理部I（cf. Fig. 5－30，Fig. 5－35）

ここでは無声摩擦音と無声破裂音の識別を行なう。

セグメントの先頭から5，母音候補カテゴリ数（NN）が0でかつBP(n)が0.1%以下であるフレームが5フレーム以上続けば無声摩擦音であると判断し，シンボル「FR」をその該当区間のフレームに割り当てる。さもなくば無声破裂音と判断し，無声破裂音検出部の「破裂音処理部」へ判断をゆだねる（cf. Fig. 5－35）。

(4). 摩擦音処理部 (cf. Fig. 5-31, Fig. 5-32, Fig. 5-33)

この処理部では，有声子音のうちの有声摩擦音 /z/ の検出と，無声摩擦音の中で有声化しやすく /z/ と混同されやすい /h/ の識別を実行する。この処理には次の2つのパラメータを使用する。

①. L_z:

HP ≧ 18 (%) かつ BP ≦ 15 (%) でさらに母音候補カテゴリ数 (NN) が2以下である連続区間の継続時間 (フレーム数).

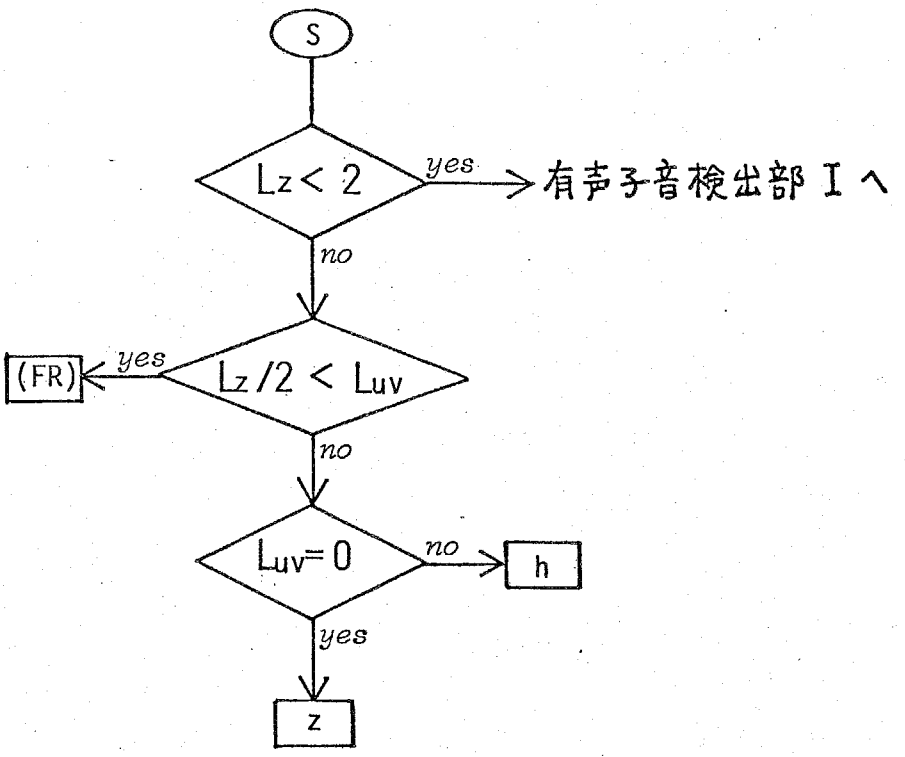
②. L_{uv}:

上記①の条件を満たす連続区間内で，PWR < 0.1 × 10^8 かつ，BP < 0.1 (%) である連続区間の継続時間 (フレーム数).

L_z, L_{uv} 2つのパラメータを使った摩擦音処理部のアルゴリズムを，Fig. 5-25 に示す。同図中 [FR], [Z], [h] 各部では，それぞれのシンボル「FR」,「Z」「H」をセグメントの先頭フレームから L_z を求める条件 (①) に適合した連続区間の終点のフレームに割り当てる。

(5). 有声子音検出部 I (cf. Fig. 5-34)

この処理部では，有声子音であるか母音単独であるかを判断する。有声子音と母音との違いは，立ち上がり部分における BP (n) の変化によく現われ，母音はなめらかに BP (n) が変化するのに対し，子音では大きなピークが現われる。そこで立ち上がり部での BP (n) のピークを検出することにより，有声子音を検出する。そのアルゴリズムを Fig. 5-26 に示す。有声子音と判定されたなら，シンボル「VC」をセグメントの先頭フレームから，PWR の極大を示すフレームの1つ手前のフレームまでに割り当てる。



Lz : HP ≧ 18, BP ≦ 15, and NN ≦ 2

Luv : HP ≧ 18, BP ≦ 0.1, NN ≦ 2, and PWR < 0.1×10^{8}

Fig. 5-25 摩擦音処理部の処理アルゴリズム.

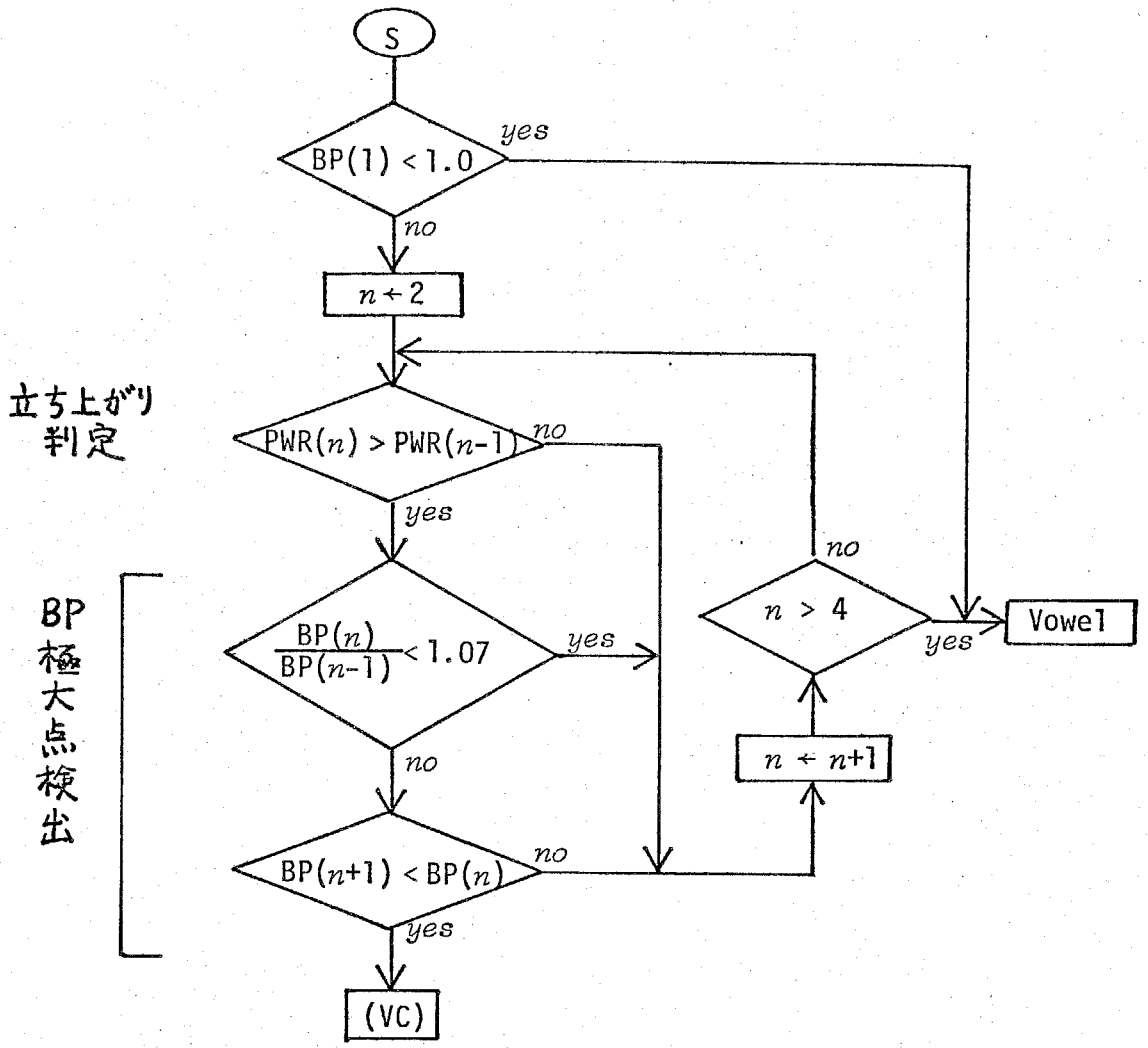


Fig. 5-26 有声子音検出部Ⅱでの処理アルゴリズム.

次の（6）～（8）の各処理部は，直前に無音部を伴わないセグメントでの子音検出処理部である。

（6）．無声子音処理部Ⅱ（cf. Fig. 5－29，Fig. 5－33）

Type 0 セグメントについては先頭フレームから，Type 0 以外のセグメントについてはサブセグメント1の先頭フレームから検索を始め，次の諸量を求め無声子音を検出する。

①．$PWR(n) < 0.1 \times 10^8$ かつ $BP(n) \leqq 0.2$（%）である連続区間を探し，次のパラメータを決める。

L_{TK} ：この連続区間の継続時間をフレーム数で表示したもの。

l_s ：この連続区間の先頭フレーム番号（セグメントの先頭から数える）.

l_e ：この連続区間の終端フレーム番号（セグメントの先頭から数える）.

$$L_{TK} = l_e - l_s + 1$$

②．①の条件を満たし，かつ，$HP(n) \geqq 10$（%），かつ，$NN = 0$ である区間を探し，次のパラメータを決める。

L_{FR} ：この連続区間の継続フレーム数．（$L_{FR} \leqq L_{TK}$）

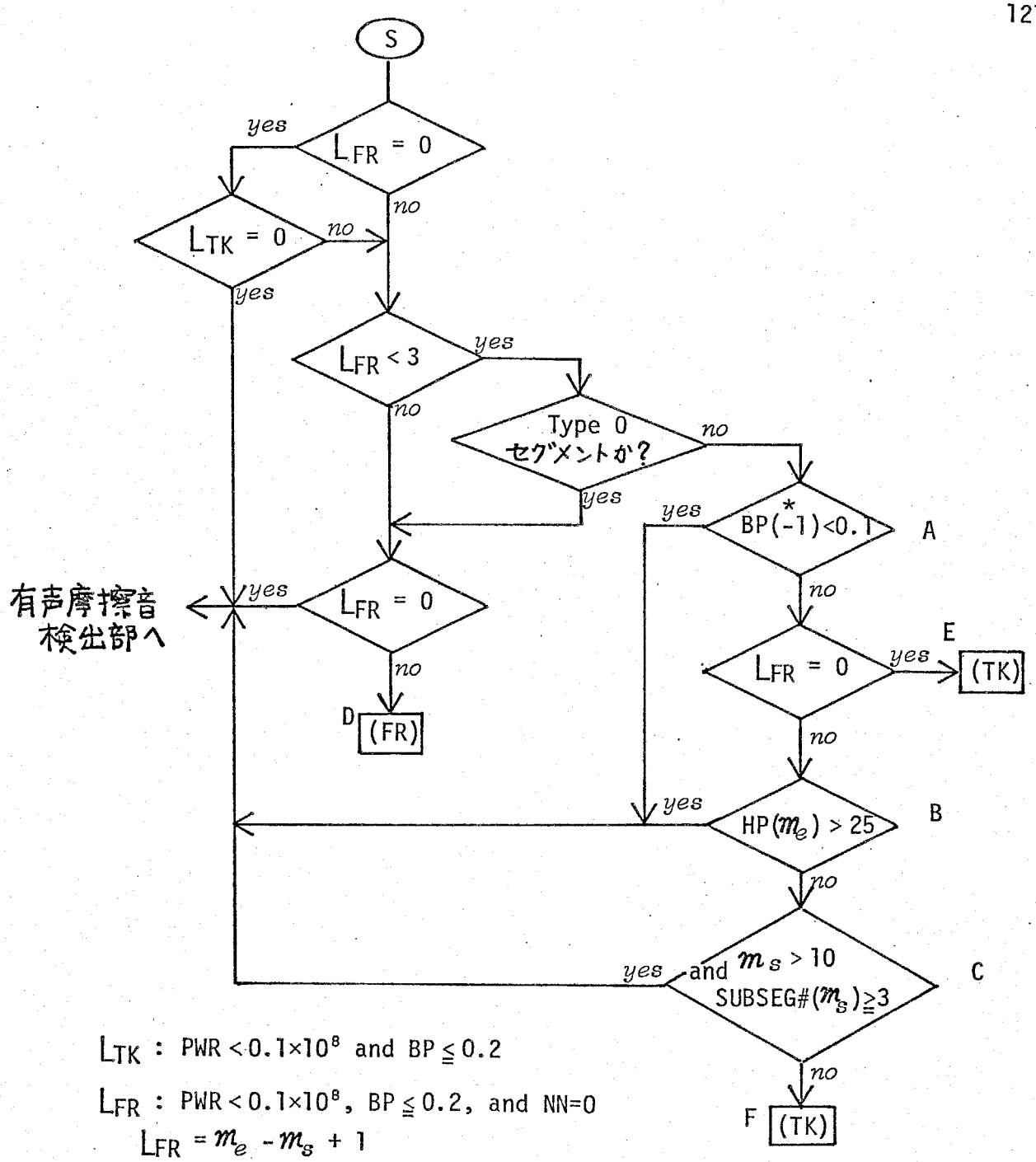
m_s ：この連続区間の先頭フレーム番号．（$m_s \geqq l_s$）

m_e ：この連続区間の終端フレーム番号．（$m_e \leqq l_e$）

$$L_{FR} = m_e - m_s + 1$$

上記①の条件は，無声子音であることの判定条件である。また②の条件は，無声摩擦音はスペクトルパターンが母音のそれとは大きく異なり（$NN = 0$），かつ高周波帯域に大きなパワーのかたよりを示すことから導入された。

Fig. 5－27に，上記のパラメータを用いた無声子音検出のアルゴリズムを示す。同図中 $BP(-1)$ は，直前のセグメントの最終フレームのBPの値



L_{TK} : PWR < 0.1×10^8 and BP ≦ 0.2

L_{FR} : PWR < 0.1×10^8, BP ≦ 0.2, and NN=0

$L_{FR} = m_e - m_s + 1$

m_s : the start frame# of the part of L_{FR}.

m_e : the end frame# of the part of L_{FR}.

* BP(-1) is the value of BP of the last frame of the preceding segment.

Fig. 5-27 無声子音処理部IIでの処理アルゴリズム.

を示す。以下，Fig．5－27中のA～Fの各部での処理について説明する。

まずAでは，直前のセグメントが無声摩擦音の子音部であった場合，それに接続する母音部の先頭が無声破裂音として検出されないように判断する。

判断Bは，摩擦音とくに/h/と無声破裂音との混乱を避けるために行なう。

判断Cは，無声破裂音はセグメンテーションの定義から各セグメントの後ろの方に出現することはないのにもかかわらず，その候補区間がセグメントの後半部に現われた場合の誤認識を防ぐためのものである。

D～Fに関しては，子音が検出された場合のそれぞれのシンボルの割り当て方法について述べる。

処理Dでは，サブセグメント1の先頭フレーム（Type 0セグメントの場合は，セグメントの先頭フレーム）から，m_eフレームまでの全フレームにわたり，シンボル「FR」を割り当てる。（cf．Fig．5－29）．

処理Eでは，$SUBSEG\#(n) \geq SUBSEG\#(l_s)$ かつ，$n \leq l_e$ なるフレーム n にシンボル「TK」を割り当てる。

処理Fでは，$m_s \leq n \leq m_e$ であるフレームnが，或いは $n < m_s$ であっても $SUBSEG\#(n) = SUBSEG\#(m_s)$（$m_s$フレームと同一サブセグメント）であるフレーム$n$に，シンボル「TK」を割り当てる。（cf．Fig．5－33）．

次の（7），（8）の処理は，有声子音を検出するための処理であるが，有声子音を含むセグメントは，少なくとも5フレーム以上続くことから，4フレーム以下の短かいセグメントに対しては処理は行なわない。

（7）．有声摩擦音検出部　（cf．Fig．5－34）

この処理部では，有声摩擦音/z/の検出と無声摩擦音/h/の検出を行なう。サブセグメント1の先頭フレームか，或いはサブセグメント0が4フレー

ム以上続く場合はセグメントの先頭から54フレーム以後から，次の2つの連続区間を検出し，/z/と/n/の検出・分類を行なう。

①. HP(n) ≧ 18(%) かつ BP(n) ≦ 15(%) かつ NN = 0である連続区間.

L_z ：この連続区間の継続フレーム数.

l_s ：この連続区間の先頭フレーム番号（セグメントの先頭から数える).

l_e ：この連続区間の終端フレーム番号（セグメントの先頭から数える).

$$L_z = l_e - l_s + 1$$

②. 上記①の区間内で BP(n) < 4(%) かつ PWR(n) < 0.1×10^8 である連続区間.

L_H ：この連続区間の継続フレーム数.

Fig. 5-28に，これらのパラメータを用いた有声摩擦音検出部での処理アルゴリズムを示す。/z/及び/n/と判定されたならば，それぞれのシンボル「Z」，「H」を，l_sフレームを含むサブセグメントの先頭からl_eフレームまでの全フレームに割り当てる。

/z/は有声子音であるために，無声摩擦音の場合のように後続母音が無声化することはなく，また子音部だけがセグメンテーションにより母音と切り離されることもないので，同一セグメント内に後続母音部が存在しないはずはないことから，Fig. 5-28中のAの判断を実行する。

また，/z/が，Type 2以上のセグメントのサブセグメント2以後に検出された場合（SUBSEG#(l_s) ≧ 2）は，そのセグメントの先頭部（サブセグメント1の先頭）には別の有声子音が存在する可能性があるので，Fig. 5-28中のBのパスを通り，「有声子音検出部II」へ進み有声子音検出処理を実行する。

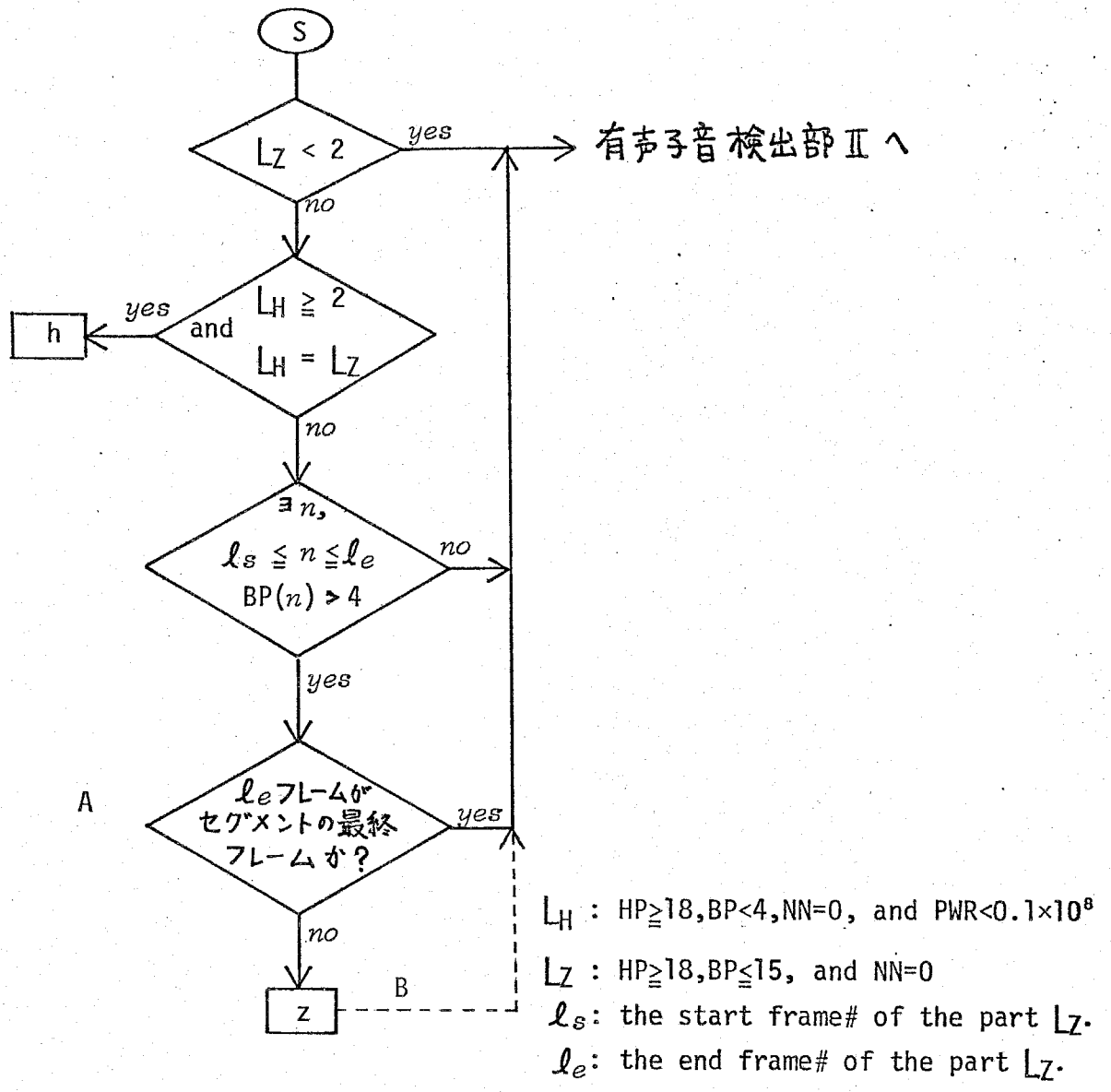


Fig. 5-28 有声摩擦音検出部の処理アルゴリズム．
SUBSEG#(ℓ_s) ≧2の場合，パスBを通り，「有声子音検出部Ⅱ」へ進む。

（8）．有声子音検出部Ⅱ （cf. Fig.5－30，Fig.5－31，Fig.5－33）

有声子音はType 0 のセグメントには存在しないことから，Type 0 以外のセグメントについて子音検出を実行する。子音検出は，「有声子音検出部Ⅰ」の処理で述べたようにパラメータBP（n）の極大点を検出することにより行なう。検出区間は，サブセグメント1の直前の2フレームと，サブセグメント1の先頭から5フレーム以内で，パワーが増加傾向にある区間である。

BP（n）が次の2条件のどちらかを満たせば，有声子音と判断する。

①．$BP(n) \geqq 19$（%）で

$$\frac{BP(n)}{BP(n-1)} > 1.0 \quad かつ \quad \frac{BP(n+1)}{BP(n)} < 1.0$$

②．7（%）$< BP(n) < 19$（%）で

$$\frac{BP(n)}{BP(n-1)} > 1.07 \quad かつ \quad \frac{BP(n+1)}{BP(n)} < 1.0$$

有声子音と判断されれば，サブセグメント1の先頭から，サブセグメント1内でパワーが増加しているフレーム（$PWR(n) < PWR(n+1)$ なるフレーム n）に シンボル「VC」を割り当てる。

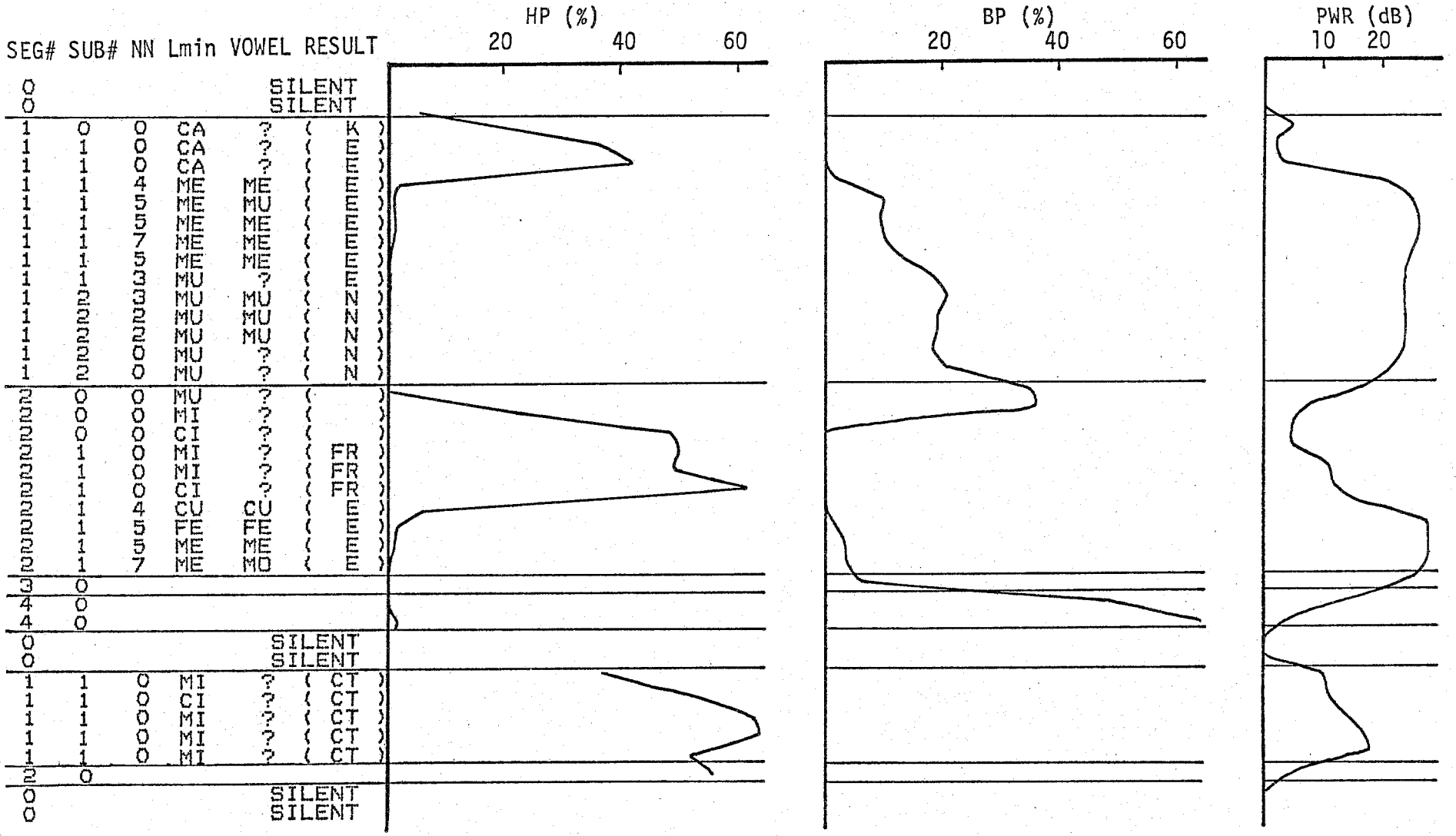


Fig. 5-29 「建設」の認識結果.

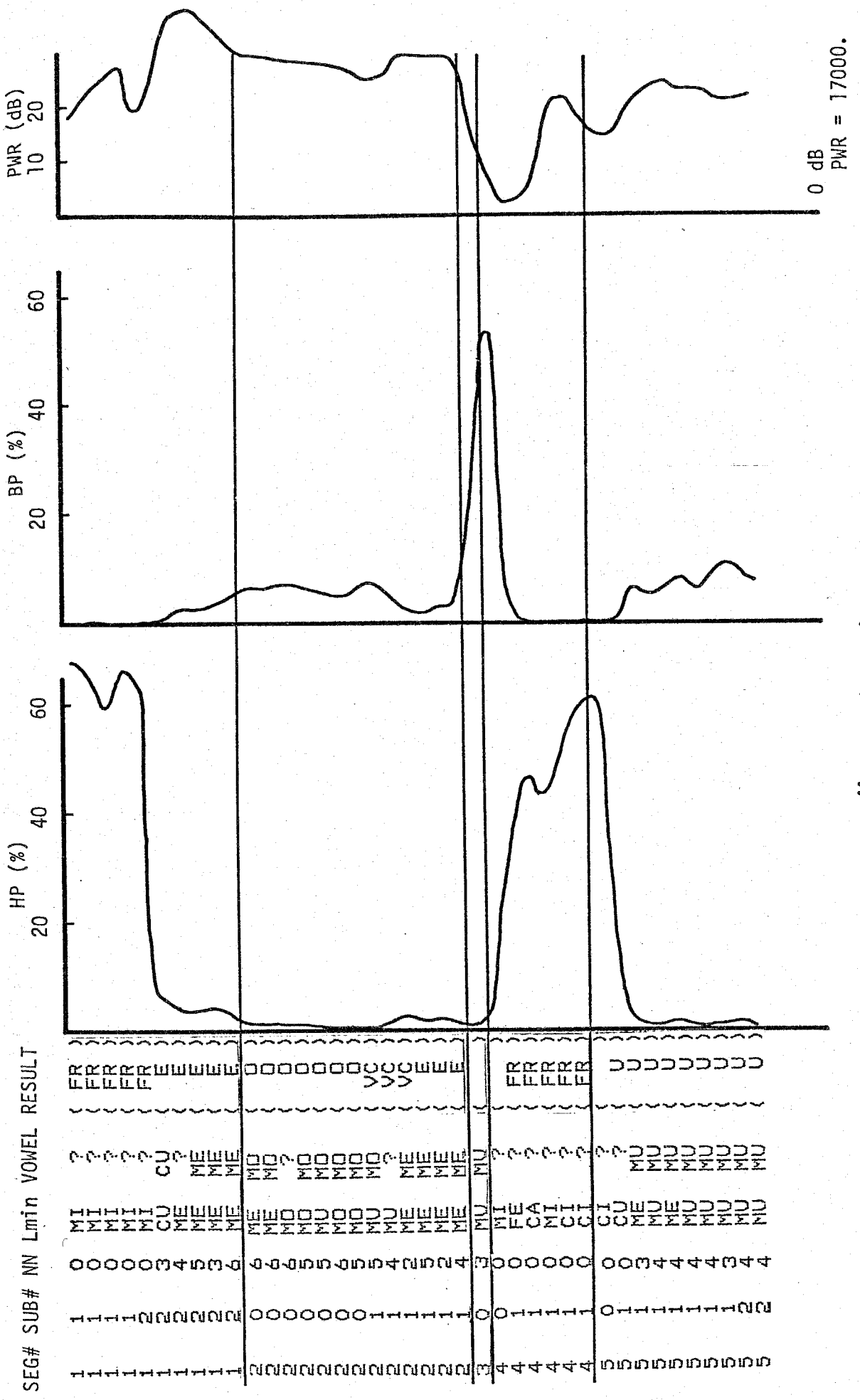


Fig. 5-30 「先月」の認識結果

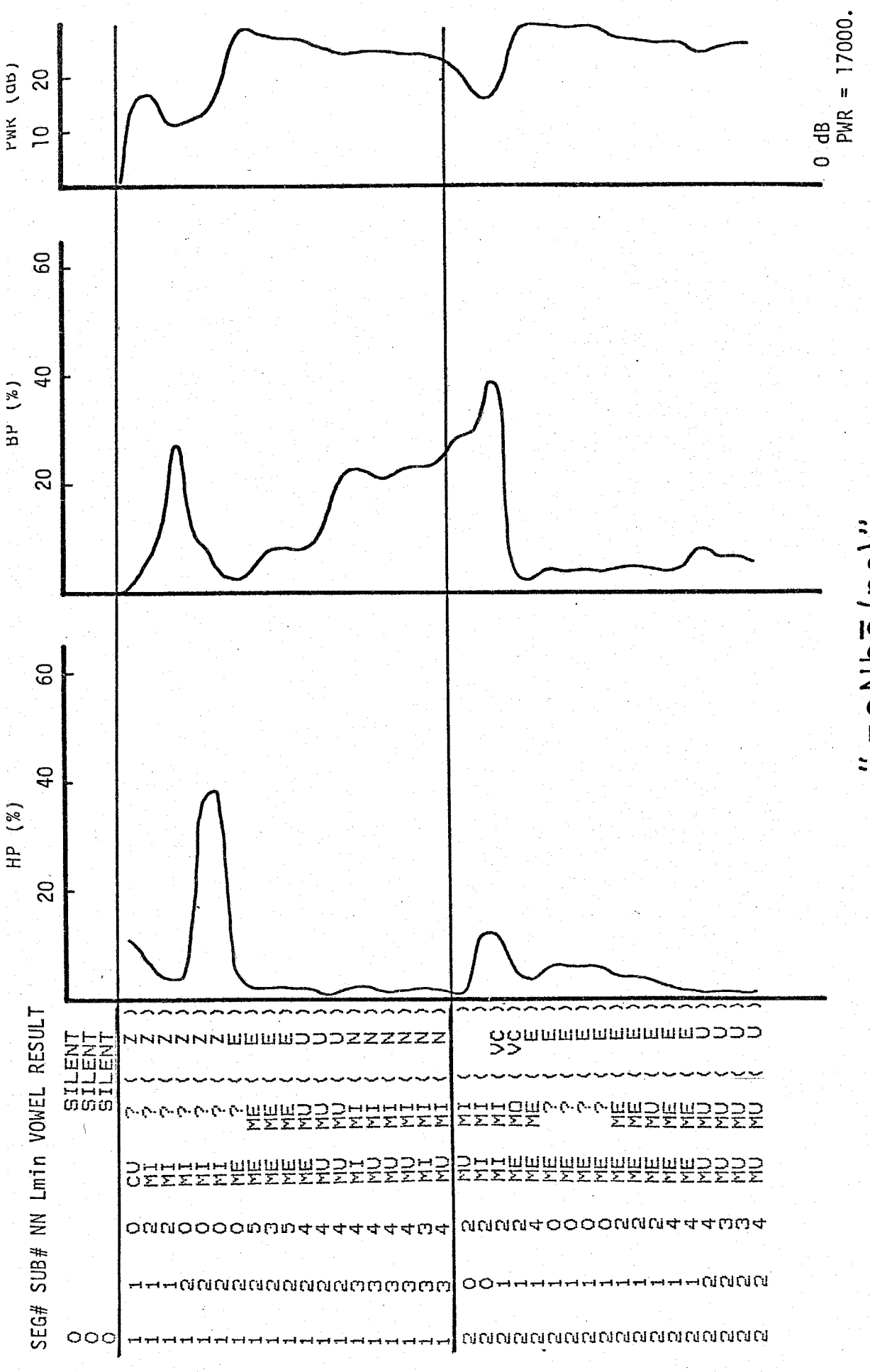


Fig. 5-31 「全米(の)」の認識結果.

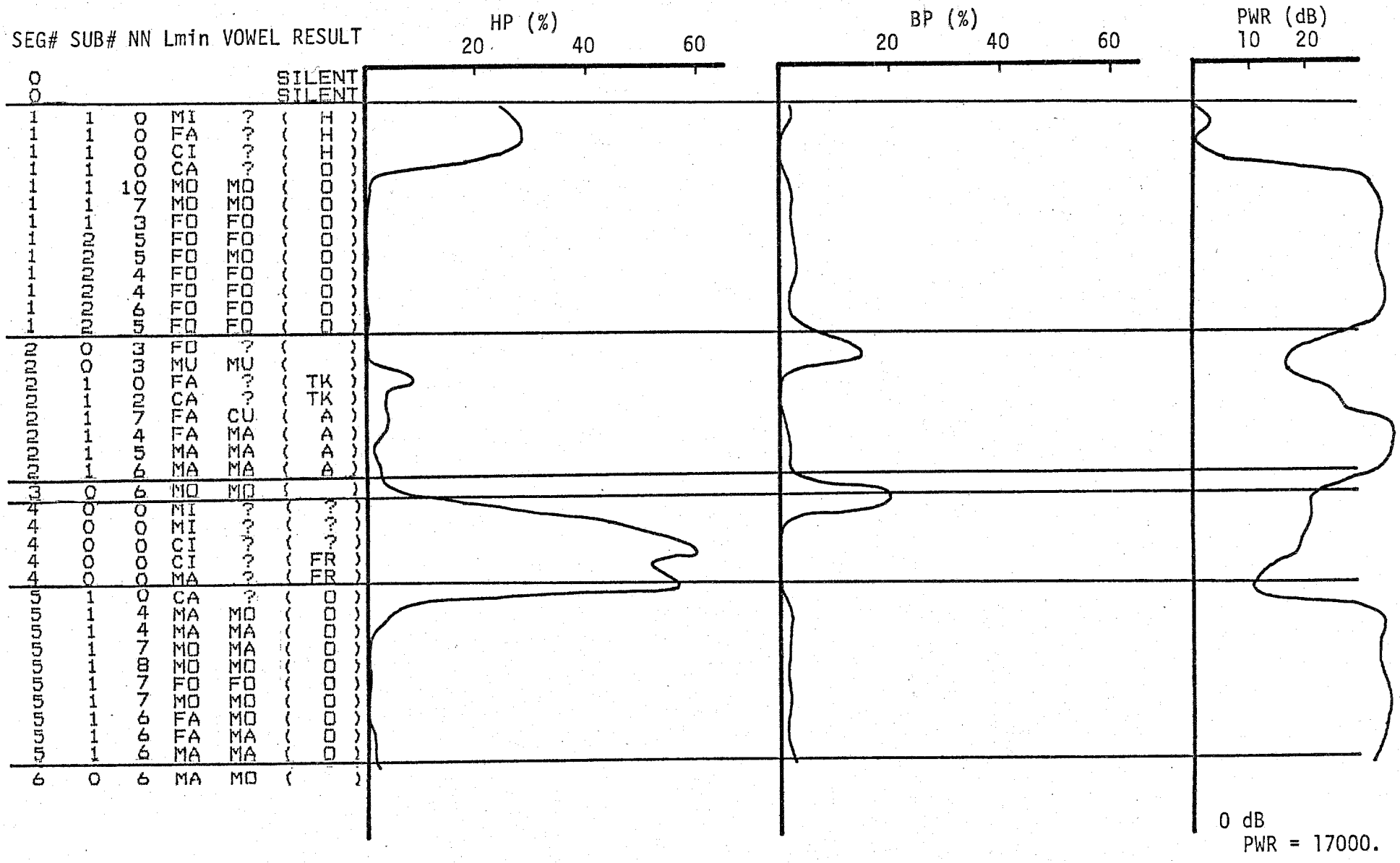


Fig. 5-32 「放火騒（ぎ）」の認識結果.

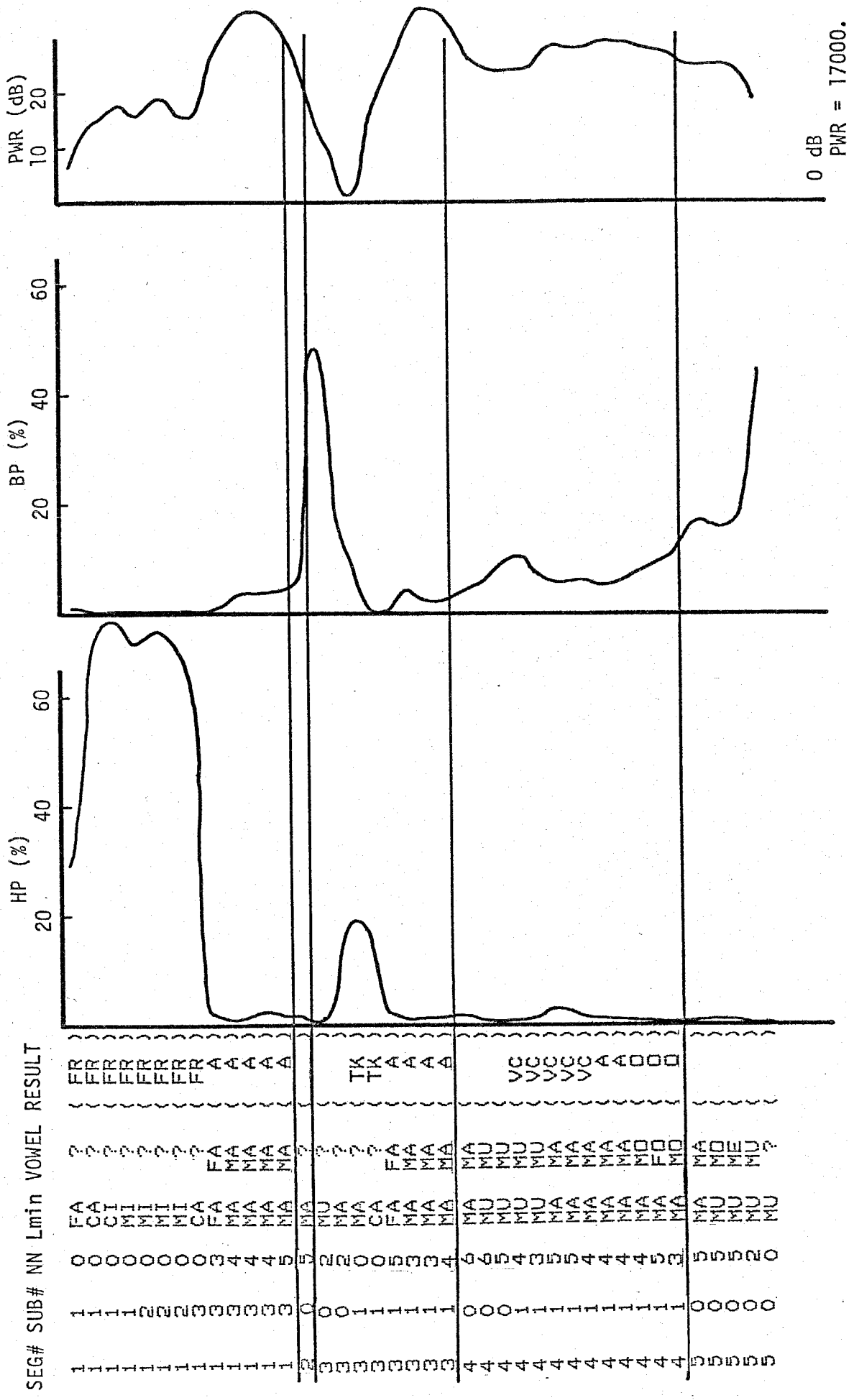


Fig. 5-33 「さかなを」の認識結果.

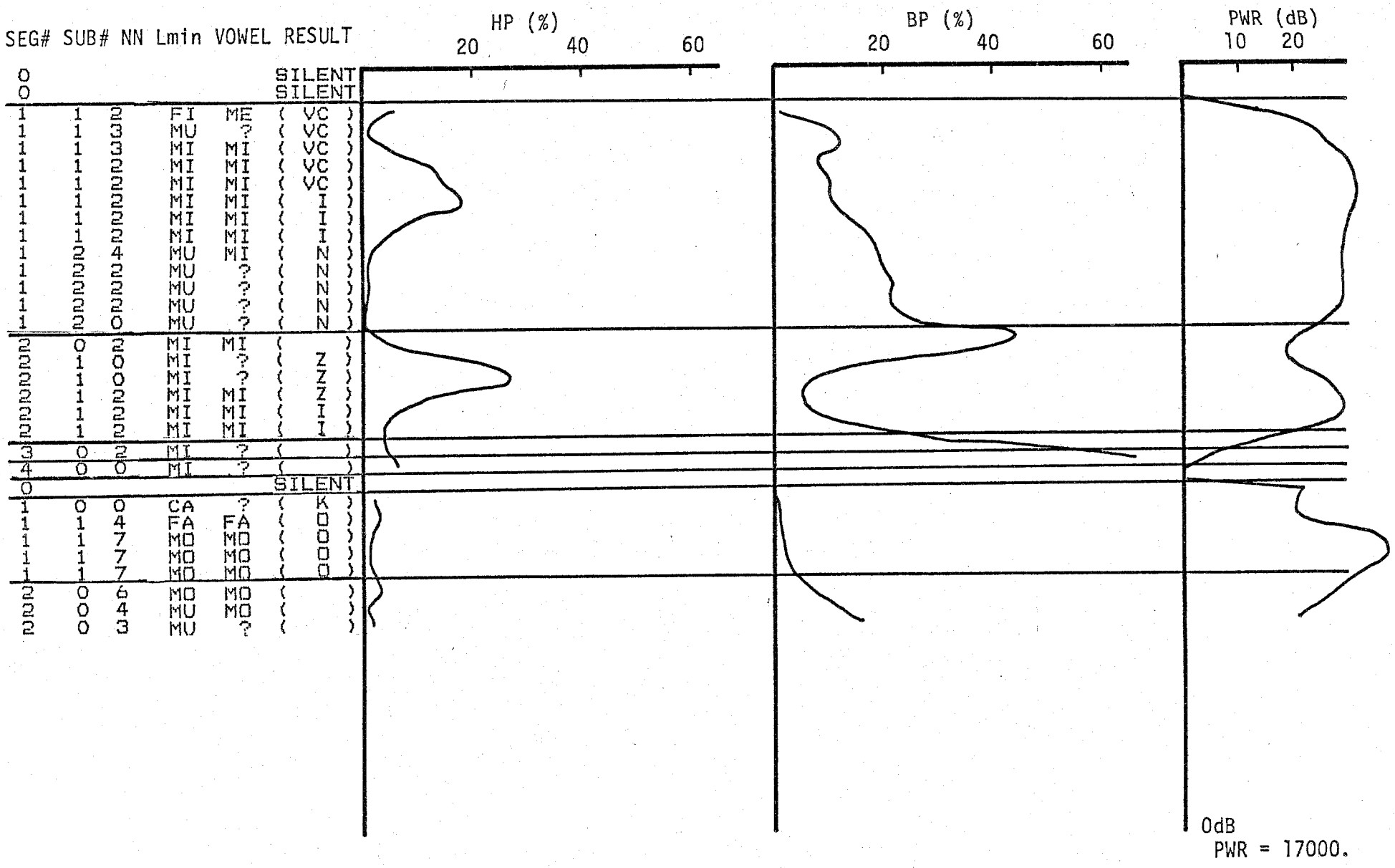


Fig. 5-34 「臨時国(会)」の認識結果.

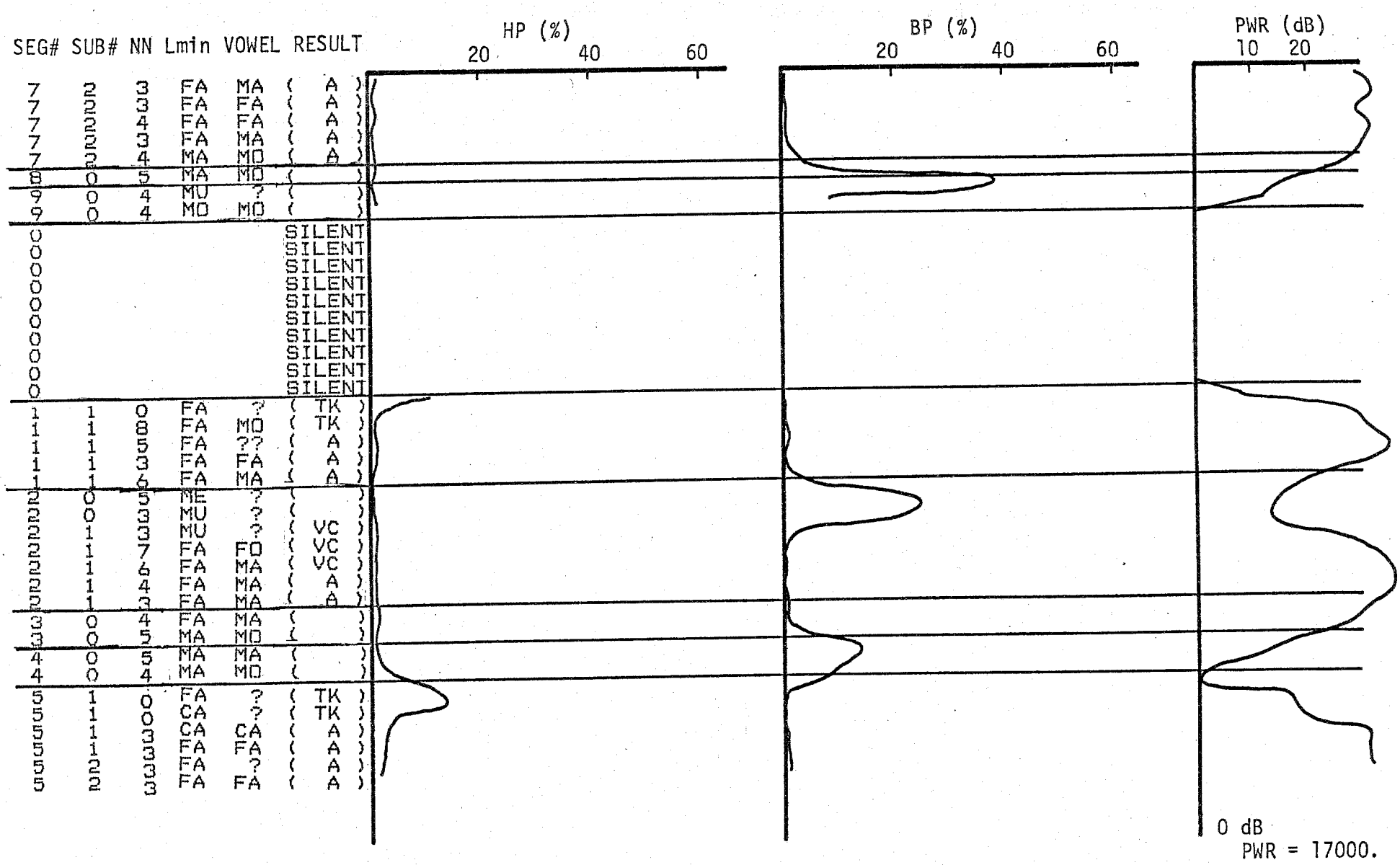


Fig. 5-35 「あ，たばか（り）」の認識結果.

5．3．3　母音部の再検討．

母音の認識，子音の検出・分類の各操作を経た後，以下の処理を行ない母音部分の再検討を行なう。この再検討は，Type 0以外のセグメントに対して実行する。Type 0 セグメントで，母音とも子音とも判断されなかったものはわたりの部分とみなす。

（1）．1つのサブセグメント内でシンボル「？」が2フレーム以上続いている区間があれば，その該当区間内で母音カテゴリ系列を使い音韻クラス（/a/～/o/の5クラス，非母音クラス/?/は除く）間の多数決をとり，母音を決定する。同票のクラスが存在したり，母音カテゴリ系列がすべて/?/である場合は，距離最小系列を用い多数決をとる。その結果，母音クラスのシンボルを該当フレームに割り当てる。

（2）．1つのサブセグメント内で，母音カテゴリ系列では3フレーム以上同一母音クラスが持続しているにもかかわらず，母音認識結果（多数決の結果）がその母音クラスと異なる場合，その区間がサブセグメントの終端フレームを含む場合に限り，同一母音クラスが持続している区間の母音シンボルを，その母音クラスのシンボルに書き換える（cf. Fig. 5－33 "sakanao"）。また，母音カテゴリ系列で2フレーム同一母音クラスが続き，かつ直前のフレームの母音カテゴリ系列のシンボルが「？」である場合も，同様に取り扱う。Fig. 5－36 にこの一例を示す。

SUBSEG#	母音カテゴリ	母音認識結果	再検討後
1	MA	(A)	(A)
1	MA	(A)	(A)
1	MA	(A)	(A)
1	?	(A)	(O)
1	FO	(A)	(O)
1	FO	(A)	(O)
2	?	(I)	(I)
2	MI	(I)	(I)

Fig. 5-36　母音部再検討（2）の一例．

5．4 認識実験．

NHK放送のニュースから録音した男性8名，女性5名の連続音声データに対して，認識実験を試みた。データの長さは，1例を除き 3．2768 sec（256フレーム）とした。以下，全データのセグメント結果（サブセグメントは省略）及び認識結果を示す。また，Fig．5－37には，1例について実際の音声波形との対応を示す。

下記の認識結果中の Speaker class [M]，[F]，[C] は，母音認識から得られた話者の属性クラスであり，これによりBP（n）を算出するチャネル区間が決まる。また認識結果中のシンボルは，それぞれ以下の意味を表わす。

(VC): N, n, m, r, b, d, g, w, y　(TK): p, t, k　(CT): ch, ts
(FR): s, sh, ch, ts, h　x : undetected.

* unvoiced vowel

Speaker M1 (male)　「これを受けて衆議院の議院運営委員会は・・・」
Speaker class[M]

segmentation

Result	ko	xeo	u	ke	(TK)e	(FR)i	uzi	i	N(VC)o	(VC)i	iN	u	(VC)ueiui	i	N	ka	i(VC)a								
Input	ko	reo	u	ke	t e	sh y	ugi	i	N n o	g i	iN	u	N ei i	i	N	ka	i w a								

Speaker M1　「・・(は)，日程等について協議した結果，臨時(国)…」
Speaker class [M]

Result	a	(VC)i	(TK)e	(VC)a	(VC)a	(VC)i	(CT)	i	(TK)e	ku	u	xi	(FR)	(TK)a	ke										
Input	a	n i	t e	n a	d o	n i	tsu*	i	t e	kyou	gi	shi*	t a	ke											
	ka	(VC)iN	zi	ko																					
	ka	r iN	zi	ko																					

Speaker M2　「・・・いには，小松辰雄が全米の攻撃を3に(ん)…」
Speaker class [M]

Result	i	(VC)iua	(TK)o	(VC)a	(CT)	(TK)a	(CT)o	eia	zeN	(VC)euo	(TK)o	(VC)i	e												
Input	i	n iwa	k o	m a	tsu*	t a	tsu*o	ga	zeN	b eno	k o	g	e												
	kio	(FR)	a	(VC)i																					
	kio	s	a	N ni																					

Speaker M3　　「神奈川県相模原市の道路の建設工事現・・・」

Speaker class [M]

Result	ka	(VC)a	(VC)a	xa	(TK)e	u(VC)u	(FR)	a	xa	(VC)i	(VC)a	xa	(FR)	uo	u
Input	ka	n a	g a	wa	k e	N	s	a	ga	m i	h a	ra	shi*	no	

	xo	(VC)o	(VC)eo	keN	(FR)e	(CT)	kuo	(VC)i	(VC)e	N
	do	r o	n o	keN	s e	tsu*	k o	z i	g e	N

Speaker M4　　「韓国の金大中氏が、今日、ソウルの自宅で・・・」

Speaker class [M]

Result	(TK)a	(VC)u	ko	kuxo	(CT)i	N(VC)u	(VC)a	i	(CT)e	(FR)	(VC)a	(CT)uu
Input	k a	N	ko	kuno	k i	N	d a	i	chyu	shi*	g a	kyou

	(FR)	o	(VC)e	(VC)auzi	(TK)a	(TK)u	(VC)e
	s	o	r u	n o zi	t a	k u	d e

Speaker M5　　「さかなを食い荒らすイルカの被害に悩まされている（長）・・・」

Speaker class [M]

Result	(FR)a	(TK)a	(VC)ao	kiia	(VC)a	(FR)	u	N	(TK)i	(VC)u	(TK)a	(VC)x
Input	s a	k a	n ao	kuia	r a	s	u		i	r u	k a	n o

	hx	(VC)a	x(VC)xxaxa	(VC)a	(FR)	a	(VC)e	kei	hu	(VC)a
	hi	g a	i n inaya	m a	s	a	r e	tei	ru	n a

Speaker M6　　「先月、放火騒ぎがあったばかりの東京池袋・・・」

Speaker class [M]

Result	(FR)e	o(VC)e	(FR)	u	ho	(TK)a	(FR)	o	(VC)e	(VC)aa	(TK)a	(VC)a	(TK)a
Input	s e	N g e	ts	u	ho	k a	s	awa	g i	g aa	t a	b a	k a

	(VC)io	o	(TK)o	kioi	(TK)e	(VC)x	(TK)o	xo
	r in	o	t o	kyoi	k e	b u	k u	ro

Speaker M7　　「1988年のオリンピックに名古屋市が立候補(す)…」
Speaker class[F]

Result	h	e	(VC)i	k̄ e	(VC)a	(TK)xa	(TK)i	zo	(VC)a	(FR)	i	(VC)iexxxo	(VC)i	i
Input	s	e	N	kyu	hya	ku*ha	ch i	zyu	h a	ch	i	n eNnoo	r i	N

(TK)i	kexi	(VC)a	(VC)x e	(FR)	xa	(VC)i	(TK)o	(VC)o	(FR)
p i	kuni	n a	g oya	shi*	ga	r i	k o	h o	s

Speaker M8　　「埼玉県がこれまでおよそ100(人)…」
Speaker class [M]

Result	(FR)	a	xxa	xa	u	a	xeexa	ko	xe	xa	(VC)ea	o e	(FR)	o	(TK)ia	kuii
Input	s	a	ita	ma			keNga	ko	re	ma	d e	oyo	s	o	hy a	kuni

Speaker F1 (female)　「NHK東京FM放送 J…」
Speaker class [F]

Result	exux	(CT)	(TK)e	i	ko	kio	(TK)e	(FR)uxN	ho	(FR)	ao	a	z e	i
Input	enue	chi*	k e	i	to	kyo	e	h ueM	ho	s	o		zye	i

Speaker F2　　「それから茨城県にあります中央競馬会…」
Speaker class [F]

Result	(FR)e	(VC)e	(TK)exa	(TK)i	(VC)a	(VC)a	(FR)	xe	N(VC)ia	(VC)i	(VC)ai
Input	s o	r e	k ara	i	b a	r a	gi	ke	N n ia	r i	m a

(FR)	he	(CT)	uo	kie	(VC)a	ka	x
s	u	chy	uo	ke	b a	ka	i

Speaker F3　　「栃木県，北寄りの風一時南寄り…」
Speaker class [F]

Result	ke	(FR)ixi	ke	(VC)u	(FR)	(TK)a	(VC)e	(VC)ixu	(TK)e	zei	(CT)i	(FR)
Input	to	ch igi	ke	N	ki	t a	y o	r ino	k a	zei	ch i	z

iuieaxie	(VC)i
iminamiyo	r i

Speaker F4　　「政府は，1988年のオリンピックを…」
Speaker class [F]

Result	(FR)ie	(FR)uxa	(FR)	u	k i	(FR)e	(TK)exa	(FR)i	hiu	(FR)a	(CT)i	u
Input	s e	h uwa	se	N	Kyu	hya	k uha	ch i	zyu	h a	ch i	

	(VC)exxxx	(VC)iN	(TK)i	k o
	n eNnoo	r iN	p i	ku*o

Speaker F4　　「オリンピックを誘致する場合は，関連施設や公立事業等についての地元の計画案では膨大な費用がかかるため，財政…」
Speaker class [F]
(768 フレーム，9.8304 sec)

Result	u	(VC)i	x	(TK)i	(TK)uo	(VC)iN	(CT) u	(VC)e	(VC)aaa	i	(VC)oa	(TK)aN
Input	o	r i	N	p i	k uo	y u	chi*su	r u	b awa	i	w a	k aN

	(VC)eN	h	(FR)e	(CT)ia	(TK)ou	(VC)i	(TK)u	(VC)i	xoo	(VC)ea	xe
	r eN	shi*	s e	tsu*ya	k ou	r i	ts u	z i	gyo	n a	do

	(VC)ui	(FR)i	N	(TK)xuo	zixx	(TK)uxu	(TK)e	(TK)a	(TK)xa	(VC)u	(VC)e
	n i	tsu*i		t eno	zimo	t ono	k e	k a	k ua	N	d e

	(VC)a	xuu	(VC)a	ixa	(FR)iu	o	(VC)a	(TK)a	kae	(VC)e	(TK)aN	e
	w a	bou	d a	ina	h iy	o	g a	k a	ka	r u	t am	e

	ze	(VC)e	(FR)	i
	za	i	se*	i

Speaker F5　　「今日，北日本を通った前線が，明日はまた関…」
Speaker class [C]

Result	(TK)io	xx	xe	(VC)i	(VC)uo	(VC)ou	(TK)o	u	xe	ze	N	(FR)	keuxa	a	(FR)u
Input	k yo	ki	ta	n i	h o	N o	t o		ta	ze	N	s	eNga	a	s u

	(VC)oa	(VC)a	(VC)o	koN
	w a	m a	d a	kaN

NHKMA1

これを受けて 衆議院の議院運営委員会は…

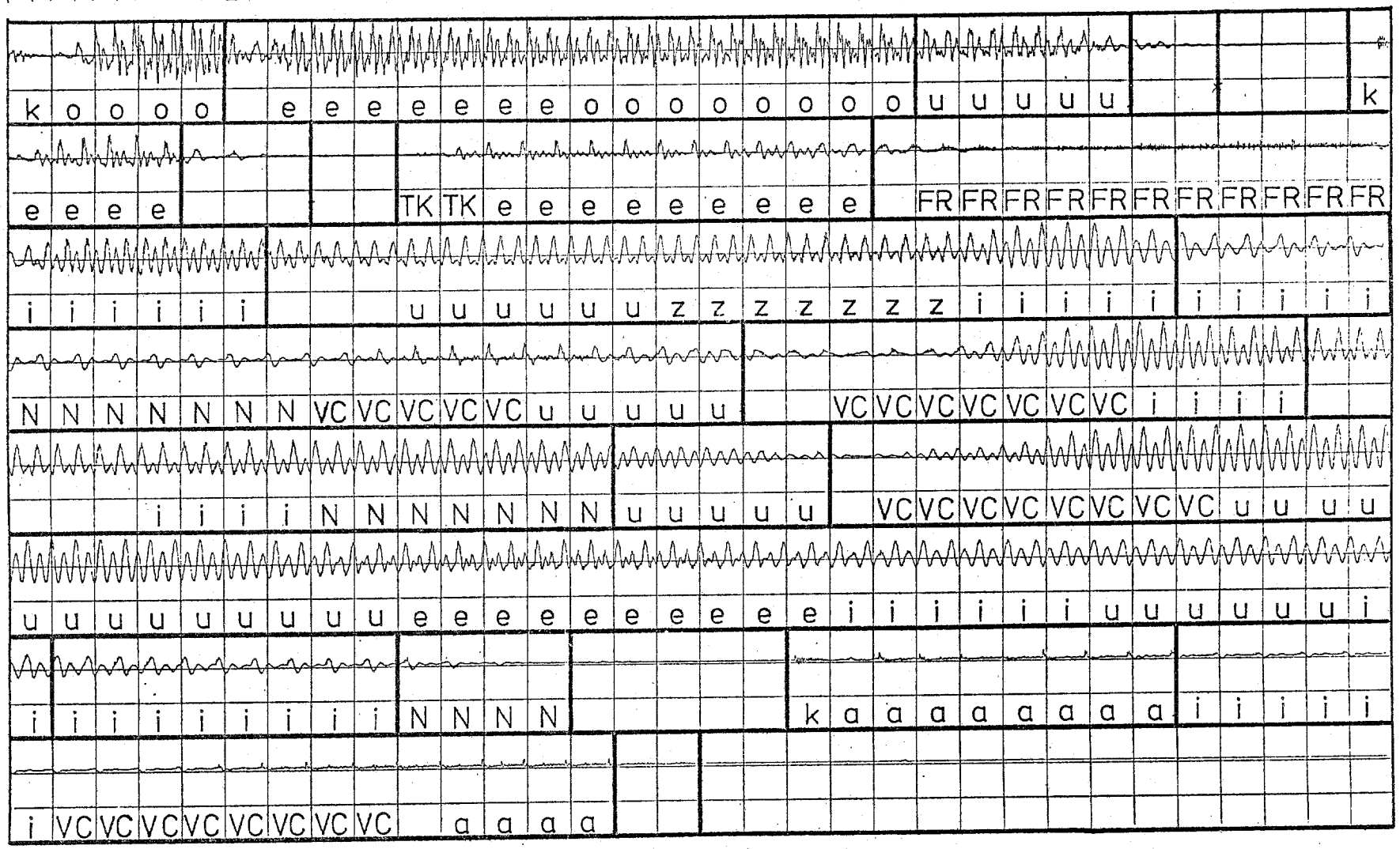


Fig. 5-37 連続音声認識結果（音声波形との対応）.

5.5 検討.

前節で示した実験結果から，無声化していない母音についてのConfusion Matrixを作成するとTable 5－1のようになる。なお，無声化していない母音の総数は315個で，無声化母音は16個である。1つの入力母音に対して母音連鎖という形で認識されたものの中にその入力母音のクラスが含まれている場合も正解と見なした場合（Table 5－1の太線枠内）の認識率は，Table 5－2のようになる。

不特定話者の連続音声認識で，母音認識率が女性を含めて90.0%という高い値が得られたということは，この母音認識システムが非常に優れたものであることを示している。単母音認識システムからの変更は，本実験では第2段階の検証アルゴリズムのわずかな修正だけであり，第1段階で用いる参照パターンに関しても検討を加え適当な修正を加えれば，さらに高い母音認識率が得られる可能性がある。この母音認識システムは2段階に分かれていて，それぞれが相補的に組み合わされているので，flexibilityが高く，連続音声認識への適応性に優れたものであることがわかる。また，サブセグメント内の母音決定には多数決という単純な手段を用いたにもかかわらず，このような高い認識率を上げることができた背景には，スペクトル分析に用いた基底膜モデルのQ値が約2と低いために，得られたスペクトルパターン上に調音結合の影響が顕著に現われなかったものと考えられる。

さて，個々の母音認識結果についてながめてみると，/u/と/o/，特に女声の/u/と/o/の認識率が低いのが目につく。この両母音について，誤認識，或いは未検出の部分について結合する子音の種類との関係を調べてみると，/o/については鼻子音，半母音，拗音と結合した場合が男声で8例中7例，女声で10例中9例と圧倒的に多いことがわかる。また/u/については，鼻子音，半母音，拗音との結合例は男声で8例中4例，女声で8例中2例と，/o/の場合ほど多くはなく，無声子音と結合した場合が男声で2例，女声で4例と多いことが目につく。これは母音/u/が無声子音と結びつくと無声化

Table 5-1　連続音声中の母音認識結果の Confusion Matrix.

IN \ OUT		/ua/ /au/	/a/	/ae/ /ea/	/e/	/ei/ /ie/	/i/	/iu/ /ui/	/u/	/uo/ /ou/	/o/	/oe/ /eo/	/oa/ /ao/	X
MALE	a		52		1						2			
	e			1	23	1								
	i	1			1		32	1						4
	u				4		1		12		2			1
	o	1	1		1				3	2	25	1		2
FEMALE	a		28	1	7				1		2			
	e				17	2								3
	i				1		28	1						3
	u				5		2		10					1
	o		1		5				4	2	12		1	3

X: undetected

Table 5-2　連続音声中の母音認識結果（認識率）.

		Total number	CORRECT single	CORRECT connected	ERROR	un-detected	SCORE (%)		
Male	/a/	55	52	0	3	0	94.5	85.7	80.0
	/e/	25	23	2	0	0	100.0		
	/i/	39	32	1	2	4	84.6		
	/u/	20	12	0	7	1	60.0		
	/o/	36	25	3	6	2	77.8		
Female	/a/	39	28	1	10	0	74.4	72.9	
	/e/	22	17	2	0	3	86.4		
	/i/	33	28	1	1	3	87.9		
	/u/	18	10	0	7	1	55.6		
	/o/	28	12	3	10	3	53.5		

しやすく，完全に無声化しなくとも非常に不安定な状態となるからであろう。母音の鼻音化，無声化の問題解決には，さらに詳しいスペクトルパターンの解析とともに，音韻情報や言語情報の活用も必要であろう。

子音の検出・分類結果に関しては，Table 5－3に半母音，拗音を除いた子音についてのConfusion Matrixを示す。識別率は同表中の太線枠内を正解と見なして計算した。また半母音についての認識結果はTable 5－4にまとめ，拗音についてはすべての認識結果をTable 5－5に上げた。この他に子音でないところを子音と検出した例が7例あった。また，セグメンテーションの誤り（子音の直前でセグメンテーションがなされていない場合）は30例あり，Table 5－3中における子音の未検出38例中の約半数と母音と誤認識された14例中のほとんとが，このセグメンテーション誤りに起因するものである。パワーの変化だけによるセグメンテーションに基づく子音の検出には限界があり，母音の認識結果（母音カテゴリ系列や距離最小系列等）をさらに積極的に利用し，総合的な判断を行ない子音を検出するシステムを考える必要がある。半母音に関しては，母音部の認識結果をも含めて約半数が，拗音については同様に3分の1が妥当な結果が得られているが，母音の中性化によると思われる誤りが多く見うけられる。

子音の検出・分類は，本研究では連続音声中の母音部検出に付随した問題として非常に簡単なシステムしか構成しなかったが，連続音声認識システムを完成させるためには，スペクトルパターンやその変化の様子の観察から，さらに子音を再分類するシステムを考え，また，促音や拗音処理の問題を解決していく必要がある。

IN \ OUT		Vowel				Voiced Consonant						Unvoiced Consonant							
												Fricative			Stop				
		o	e	i	u	(VC)i	(VC)u u(VC)u	uN N(VC)u	N	(VC)	z	h	(FR)	(CT)	(PTK)	k	X	*total*	SCORE %
VOICED	N	1	1	.	3	1	5	2	11	5							3	32	71.9
	n	1	1	2	4		1			18							9	36	52.8
	m				1				1	5							4	11	45.5
	r									20		1					5	26	76.9
	b									7							1	8	87.5
	d									6							2	8	75.0
	g				1					11	1		1				6	20	55.0
	z									2	7		1					10	70.0
UNVOICED	h									4		3	4				3	14	50.0
	s,sh											2	25	1				28	89.3
	ch											1	3	3	1			8	75.0
	ts												2	5	1			8	87.5
	k												2	1	23	23	2	51	90.2
	t														13	3	3	19	68.4
	p														3			3	100

X: undetected

Table 5-3　半母音，拗音を除いた連続音声中の子音の検出・分類結果．

Table 5-4 半母音の認識結果.

IN \ OUT	a	i	u	e	o	ia	iu	io	ua	uo	(VC)a	(VC)i	(VC)u	(VC)e	(VC)o	(VC)oa	(VC)iN
ya	1			1		1											
yu																	1
yo				2						1				1			
wa	3				1				1		2					2	

Table 5-5
拗音の認識結果.

	IN		OUT
Male	zyū	→	zo
	hya	→	(TK)ia , (VC)a
	shyū	→	(FR)iu
	chyū	→	(CT)e
	kyū	→	ke
	kyo	→	kio
	kyou	→	kuu , (CT)uu
Female	gyō	→	Xoo
	zyū	→	hiu
	zyē	→	ze
	hya	→	(FR)e
	chyū	→	(CT)u
	kyū	→	ki
	kyō	→	kio
	kyo	→	(TK)io

5．6　むすび

4章で提案した単母音認識システムを連続音声認識に応用することを試みて，簡単な子音分類も実行する連続音声認識システムを構成した結果，単母音認識システムの第2段階で用いる検証アルゴリズムを少し変更しただけで，NHKのニュース放送というかなり高速な発話の連続音声データに対して男声母音で85．7％，女声母音で92．9％，全体で80．0％という高い母音認識率を得ることが出来た。この実験により，基底膜モデルを用いたこの母音認識システムが不特定多数の話者の個人差による変動にも強く，また，調音結合等の変動にも強いシステムであることが示された。

さらに，パワーの変化を使ったセグメンテーションに基づき，母音の認識結果や高周波帯域（7～12ch）及びピッチ周辺のパワーの集中度を用いて，子音を4つの大分類クラス　無声破裂音（p，t，k），無声摩擦音（s，sh，ts，ch，h），有声摩擦音（z），その他の有声子音（N，m，n，r，b，d，g，y，w）及び，補足的な4クラス　k，h，破擦音（ts，ch），撥音N　の各クラスに検出・分類することを試みた結果，92．8％の識別率が得られた。これらのパラメータ以外に，スペクトルパターンやその変化の様子等を用いてさらに細かい子音の特徴を捕え，また，音韻に関する情報をTop-down的に利用する方法[1]等の導入を画ることにより，子音の細分類，拗音・促音処理の問題，無声化母音の取り扱い等の問題を解決していくことが，十分に高い音韻認識率が得られる不特定話者対応の連続音声認識システムを開発するために必要である。

最後に，連続音声認識システム実用化のための，処理速度の高速化に関して言及する。

現在，本システムの演算時間は，約3．3 sec（256フレーム）の音声データを処理するのに，基底膜演算時間が約30 min，母音認識時間が約200 sec（1フレーム当り1 secの処理時間で，母音認識を必要とするフレーム数が平均200フレームとして計算），子音の検出・分類その他の処理に4．6

sec 要している。この中で最も時間を要している基底膜演算は，2・3節で述べたように基底膜モデルをハードウェア化（ディジタル回路でなくとも，アナログ回路で構成し，フレーム周期毎にその平均振幅をサンプリングするというシステムでも可）することにより，実時間で処理できるので問題はない。次に母音認識であるが，15カテゴリの各々にマイクロプロセッサを割り当て並列処理を実行するシステムを構成すれば，演算時間は現在の15分の1，約13.3 secまで短縮できる。また，母音認識処理時間のほとんどすべては，参照パターンとの距離計算に費やされている。本研究で使用したミニコンピュータ（HP2113E）の浮動小数点の乗算に要する時間は 約30 μsecであり，ハードウェアの浮動小数点演算装置があれば，この5倍程度の速度で演算が実行できるので，母音認識時間は，約2.6 secとなり，実時間で演算が可能となる。現在，HP2113Eと同程度の能力を持つマイクロプロセッサはすでに市販されているので実現は可能である。よって，実時間の3倍程度の時間があれば，子音の認識まで完了するシステムの開発が可能であろう。

第6章　結言

本研究では，聴覚系における最初の情報変換器である基底膜のモデルを用い，人間の聴き分け機構（カクテルパーティ効果）について1つのモデルを構成し，また，不特定話者を対象とした母音認識システムを提案しその有効性を示した。

基底膜モデルは，従来から用いられているようなアナログ回路モデルや数学モデルでは，各種の入力に対する基底膜応答の観察や多様な処理に対応できないので，新たに計算機処理に適したディジタル回路モデルを提案した。本モデルは，全く同じ構造の2次のローパス・ディジタルフィルタを54段縦続接続するもので非常に簡単な構成となっており，プログラミング及びハードウェア構成も容易であり，聴覚系のシミュレーションや音声認識研究には非常に有効なモデルである。

人間の聴き分け機構については，聴き分けは主にピッチの違いに着目して行なわれると仮定し，聴覚系におけるパルス相関モデルを構成し，2名の話者により同時発声された混合母音からそれぞれの成分母音のスペクトルを分離するのに成功した。パルス相関モデルは，基底膜出力を半波整流した後，さらにパルス列に変換し相関処理を行なうというモデルであるが，このパルス化がスペクトルパターンの分離能力の向上に大いに役立つことが分った。

母音認識システムは，基底膜出力から計算されるスペクトルパターンを用いるもので，第1段階として入力スペクトルパターンと参照パターンとの距離を算出し，その距離をもとに一次候補を上げ，次に第2段階で入力スペクトルパターンの形態的特徴が一次候補に上げられたカテゴリの特徴と一致するかを検証し最終判断を下すという2段階の構成となっている。また，本システムは，入力音声からピッチ抽出等を実行し話者の属性（年令・性別）を決定してから音韻の識別を実行するのではなく，話者の属性と音韻性をまとめて取り扱うという特徴を持っている。成人男性32名，同女性25名，10～12才の子供10名の発声した単母音に対して，男性で100%，女性で96.8%，子供

で98.0%，総合で98.3%という高い認識率を得ることができた。さらに，本母音認識システムを組み込んだ連続音声認識システムを構成した結果，NHKのニュース放送というかなり高速な発話の連続音声データに対して，男声母音で85.7%，女声母音で72.9%，総合で80.0%という非常に高い母音認識率を上げることができ，本母音認識システムの有効性を確認することができた。同時に実行した子音の分類は，無声破裂音（p，t，k），無声摩擦音（s，sh，ch，ts，h），有声摩擦音（z），その他の有声子音（N，m，n，r，b，d，g，y，w）の大まかな4クラスと，k，h，破擦音（ch，ts）及び撥音（N）の補足的な4クラスへの分類であるが，72.8%という識別率が得られた。

本研究では，基底膜モデルの出力を処理するという方法で研究を進めたが，基底膜を単なるスペクトル分析器としてとらえず，基底膜各位置の振動波形そのものを取り扱ったために聴き分けのシミュレーションに成功したと言えよう。また，スペクトル分析の立場から見れば異常に低いQ値を持っているが，このQ値の低さゆえに，話者の年令・性別の違いによる変動や調音結合等の影響を吸収でき，高い母音認識率を得ることができたのであろう。

最後に本稿を結ぶにあたり，不特定話者連続音声認識システムを完成させるために今後検討していかねばならない問題点にいくつか上げる。まず，本研究ではほとんど触れなかった子音認識についてであるが，スペクトルパターンの形やその時間的変化の様子から特徴を把握し認識システムを構成する必要がある。その際，子音の検出には，パワーを用いたセグメンテーションの結果だけでなく，母音認識結果をも積極的に利用する必要があろう。母音認識に関しても，参照パターンや検証アルゴリズムに検討を加えたり，また子音の認識結果をも利用しさらに高い認識率を上げる努力を積み重ねる必要があろう。それに加え，音韻に関する知識をTop-down的に利用することにより音韻認識率の向上をはかる必要があろう（特に，無声化しやすい母音や拗音等に関して）。

実用化に際しては，基底膜モデルは現在のように16 kHzから30 Hzまで

の9オクターブをカバーするものは必要なく，100Hzから56.4kHzの6オクターブ（36チャネル）もあれば十分である。また，パルス相関法を組み込み，耐雑音性能を向上させれば，優れた音声認識装置が完成するであろう。

謝辞

本研究を進めるにあたり，常に御指導下さいました河原田弘助教授に，心から感謝致します。また，貴重な助言や討論をしていただきました池辺潤教授，並びに太田道男（現・筑波大学助教授），小杉幸夫両博士にも心から感謝致します。日頃実験に御協力下さいました池辺・河原田研究室の皆様にも深く感謝致します。特に，本論文をまとめるに際し，労を厭わず御協力下さった原田哲也，神宮司誠，石川徹の3氏には感謝致します。

最後に，音声資料を提供して下さいました多くの方々にも感謝の意を表したいと思います。

参考文献

(1) 中津，好田："会話音声の機械認識における音響処理"，信学論(D)，61-D，4，p.261 (昭53-04)

(2) 三輪，新津，牧野，城戸："音声スペクトルの概略形とその動特性を利用した単語音声認識システム"，日本音響学会誌，34，3，p.186 (昭53-03)

(3) 中島，鈴木，三国："調音特徴による母音体系化の試み"，日本音響学会講演論文集，p.181 (昭53-05)

(4) 鹿野，好田："会話音声の機械認識における言語処理"，信学論(D)，61-D，4，p.253 (昭53-04)

(5) 新津，三輪，牧野，城戸："単語音声自動認識における言語情報の一利用法"，信学論(D)，62-D，1，p.24 (昭54-01)

(6) 例えば 境久雄編著，中山剛共著：聴覚と音響心理，p.183，コロナ社 (昭53)

(7) 新美康永：音声認識，共立出版 (昭54)

(8) G.von Békésy : Experiments in Hearing, McGraw-Hill (1960)

(9) W.S.Rhode : "Observations of the vibration of the basilar membrane in squirrel monkeys using Mössbauer technique", J.A.S.A., 49, 4, p.1218 (1971)

(10) B.M.Johnstone, K.J.Taylor and A.J.Boyle : "Mechanics of the Guinea pig cochlea", J.A.S.A., 47, 2, p.504 (1970)

(11) W.S.Rhode and L.Robles : "Evidence from Mössbauer experiments for nonlinear vibration in the cochlea", J.A.S.A., 55, 3, p.588 (1974)

(12) R.R.Pfeiffer and D.O.Kim : "Cochlear nerve fiber responses; Distribution along the cochlear partition", J.A.S.A., 58, 4, p.867 (1975)

(13) 渡辺武：“聴器の神経生理”，音声情報処理（比企静雄編），p.145，東大出版会（昭48）

(14) T.J.Moore and J.L.Cashin Jr. : "Response of cochlear-neucleus neurons to synthetic speech", J.A.S.A., 59, 6, p.1443 (1976)

(15) 橋本享：“聴神経系における音声の特徴パラメータの識別”，日本音響学会聴覚研究会資料，H-61-3（昭54-06）

(16) 境，氏原：“聴神経系のシミュレーション”，音声情報処理（比企静雄編），p.160，東大出版会（昭48）

(17) W.F.Cardwell : "Recognition of sounds by cochlear patterns", IEEE. Trans. MIL-7, P.179 (1963)

(18) T.B.Martin and J.J.Talavage : "Application of neural logic to speech analysis and recognition", IEEE. Trans. MIL-7, p.189 (1963)

(19) 松岡，城戸：“音声スペクトルのローカルピークの静特性のもつ音韻情報に関する検討”，日本音響学会誌，32，1，p.12（昭51-01）

(20) J.L.Flanagan : "Models for approximating basilar membrane displacement", Bell Syst. Tech. J., 39, p.1163 (1960)

(21) 境久雄：“基底膜の回路モデルとての応答”，日本音響学会講演論文集，p.39（昭40-10）

(22) D.H.Klatt and G.E.Peterson : "Reexamination of a model of the cochlea", J.A.S.A., 40, 1, p.54 (1966)

(23) B.P.Bogert : "A network to represent the inner ear", Bell Laboratories Record, 28, p.481 (1950)

(24) R.Oetinger und H.Hauser : "Ein elektrischen Kettenleiter zur Untersuchung der Mechanischen Schwingungsvorgänge im Innenohr", Acustica, 11, p.161 (1961)

(25) 大野，香田：〝基底膜の変位に関する分布定数回路的モデル〟，信学論（D），57－D，8，p．463 （昭49－08）

(26) 関口裕：〝音声情報処理システムの構成〟，東工大修士論文，（昭54－02）

(27) 河原田，亀井，関口：〝基底膜モデルを用いた音声解析装置の試作と子音の特徴観察〟，信学論（A），63－A，2，p.106（昭55-02）

(28) 鈴木秀人：〝ピッチの違いを利用した音声と音声の分離〟，日本音響学会講演論文集，p．339 （昭51－10）

(29) M.Yanagida, O.Kakusho and D.Graham-Stuart : "An answer to the Cochtail Party Problem", 10th ICA, Sydney (1980)

(30) 藤井，森田，真鍋：〝聴覚系におけるピッチ検出機構のモデル〟，信学論（D），60－D，4，p．291 （昭52－04）

(31) 井出，山田，牧野，城戸：〝Q≒6のフィルタによる女声母音の認識〟，日本音響学会講演論文集，p．545 （昭55－05）

(32) 松井，白井：〝調音パラメータによる母音識別に対する適応化の効果〟，日本音響学会講演論文集，p．547 （昭55－05）

(33) The Bipolar Digital Integrated Circuits Data Book, 1st ed., p.I 7-391, テキサスインスツルメンツ アジアリミテッド社 (1976)

発表文献

論文

(1) H.Kawarada, H.Kamei and T.Nakanishi : "A digital filter model of basilar membrane", Bull. P.M.E. (T.I.T.), 41, p.45, (March 1978)

(2) 河原田，亀井，中西：“基底膜のディジタル回路モデル”，信学論(D)，61－D，4，p.237 (昭53－04)

(3) 河原田，亀井，関口：“基底膜モデルを用いた音声解析装置の試作と子音の特徴観察”，信学論（A），63－A，2，p.106 (昭55-02)

(4) 河原田，亀井，秦：“複数話者により発声された混合音声の聴き分け機構”，信学論（A），63－A，8，p.469 (昭55－08)

(5) 亀井，河原田，後藤，原田：“不特定話者の母音認識”，電子通信学会投稿予定

学会・研究会

(1) 河原田，亀井，中西：“基底膜のディジタル回路モデル”，信学技報，PRL 77-9 (昭52－05)

(2) 河原田，亀井，溝口：“基底膜モデルによる音声の特徴抽出”，信学技報，MBE 77－58 (昭53－02)

(3) 河原田，亀井，溝口：“音声入力に対する基底膜モデルの応答”，日本音響学会聴覚研究会資料，H51－8 (昭53－06)

(4) 河原田，亀井，武田：“ピッチ，コンビネーショントーンの知覚機構”，日本音響学会聴覚研究会資料，H61－1 (昭54－06)

(5) 河原田，亀井，秦：“複数話者により発声された混合音声の聴き分け機構”，日本音響学会聴覚研究会資料，H61－2 (昭54－06)

(6) 河原田，亀井：“基底膜モデルによる音声解析”，日本音響学会講演論文集，p.543 (昭54－06)

付　録

A. 基底膜モデルの単位フィルタの係数決定式の導出.

$$H(z) = \frac{a_3}{1 - a_1 z^{-1} - a_2 z^{-2}} \qquad (A-1)$$

より, $|H(e^{j\omega T})|$ ($z = e^{j\omega T}$, Tはサンプリング間隔) は, (A−2) 式の様になる。

$$|H(e^{j\omega T})| = \left[\frac{a_3^2}{1 + a_1^2 + a_2^2 - 2a_1 \cos\omega T - 2a_2 \cos 2\omega T + 2a_1 a_2 \cos\omega T} \right]^{1/2} \qquad (A-2)$$

$$|H(1)| = g_0, \quad |H(e^{j\omega_0 T})| = g_0 \cdot g, \quad \left. \frac{d|H(e^{j\omega T})|}{d(\omega T)} \right|_{\omega=\omega_0} = 0$$

の3条件より, 次の連立方程式が得られる。

$$\begin{cases} a_3 = g_0 (1 - a_1 - a_2) & (A-3) \\ a_3^2 = g_0^2 g^2 (1 + a_1^2 + a_2^2 - 2a_1 \cos\omega_0 T - 2a_2 \cos 2\omega_0 T + 2a_1 a_2 \cos\omega_0 T) & (A-4) \\ a_1 + 4a_2 \cos\omega_0 T - a_1 a_2 = 0 & (A-5) \end{cases}$$

(A−5) 式より　$a_1 - a_1 a_2 = -4a_2 \cos\omega_0 T$　(A−6)

(A−3), (A−6) 式を (A−4) 式に代入し整理すると

$$a_1^2 + a_2^2 + 2a_2 \frac{g^2 \cos 2\omega_0 T + 2g^2 + 1 - 4\cos\omega_0 T}{g^2 - 1} + 1 = 0 \qquad (A-7)$$

が得られる。

ここで $$Q' = \frac{g^2(2P^2-1)+2g^2+1-4P}{g^2-1} \qquad (A-8)$$

$$P = \cos\omega_0 T \qquad (A-9)$$

とおくと（A－7）式は（A－10）式となる。

$$a_1^2 + a_2^2 + 2Q'a_2 + 1 = 0 \qquad (A-10)$$

ここで $x = 1/a_1$, $y = 1/a_2$ と変数変換を行ない，（A－5）式から

$$y = 1 - 4Px \qquad (A-11)$$

を得，（A－10）式に（A－11）式を代入し整理すると（A－12）式が得られる。

$$16P^2x^4 - 8PQx^3 + (2Q+16P^2)x^2 - 8Px + 1 = 0 \qquad (A-12)$$

ここで $$Q = Q'+1 = \frac{2P^2g^2+2g^2-4P}{g^2-1} \qquad (A-13)$$

（A－12）式は，さらに次のように因数分解できる。

$$(Ax^2 - 4Px + 1)(A^*x^2 - 4Px + 1) = 0 \qquad (A-14)$$

ここで A, A^* は $t^2 - 2Qt + 16P^2 = 0$ の根

$$A = Q + \sqrt{Q^2 - 16P^2}$$

$$A^* = Q - \sqrt{Q^2 - 16P^2}$$

（A－14）式から得られる2つの方程式のうち，$Ax^2 - 4Px + 1 = 0$ は，実根を持たない。

そこで　　$A^* x^2 - 4Px + 1 = 0$

変数 a_1 による表現に戻すと

$$a_1^2 - 4Pa_1 + A^* = 0 \qquad (A-15)$$

$$\therefore \quad a_1 = 2P \pm \sqrt{4P^2 - A^*} \qquad (A-16)$$

基底膜モデルには，(A－16) 式の根号の前の符号は (－) を採用する。
そして，(A－5) 式から a_2 を，(A－3) 式から a_3 を求めると以下のようになる。

$$a_1 = 2P - \sqrt{4P^2 - A^*}$$

$$a_2 = \frac{a_1}{a_1 - 4P}$$

$$a_3 = g_0(1 - a_1 - a_2)$$

ここで

$$A^* = Q - \sqrt{Q^2 - 16P^2}$$

$$Q = \frac{2P^2g^2 + 2g^2 - 4P}{g^2 - 1}$$

$$P = \cos\omega_0 T$$

次に，角周波数 ω_1 から ω_0 を求める式を導出する。 $|H(e^{j\omega_1 T})| = 1$ より (A－2) 式から，(A－17) 式が得られる。

$$a_3^2 = 1 + a_1^2 + a_2^2 - 2a_1\cos\omega_1 T - 2a_2\cos 2\omega_1 T + 2a_1a_2\cos\omega_1 T \qquad (A-17)$$

(A－17) 式に (A－3) 式を代入し整理すると

$$(1-g_0)^2(1+a_1^2+a_2^2) - 2(a_1 - a_1a_2)\cos\omega_1 T - 2a_2\cos 2\omega_1 T$$
$$= g_0^2(-2(a_1 - a_1a_2) - 2a_2) \qquad (A-18)$$

(A−18)式に，(A−6)，(A−10)式を代入し，さらに $R=\cos\omega_1 T$ と置いて整理すると，(A−19)式が得られる。

$$2R^2-4PR+(1-g_0^2)Q'-1-g_0^2(1-4P)=0 \qquad (A-19)$$

Q' に (A−8) 式を代入し整理すると (A−20) 式となる。

$$(g^2-1)R^2-2P(g^2-1)R+(1-g_0^2g^2)+2P(g_0^2g^2-1)+P^2g^2(1-g_0^2)=0 \qquad (A-20)$$

(A−20)式は，Rに関する2次方程式であるので，根を求めることができる。

$$R=\frac{P(g^2-1)\pm(P-1)\sqrt{(g_0^2g^2-1)(g^2-1)}}{g^2-1} \qquad (A-21)$$

(A−21)式で，根号の前の符号は(+)を採用し，Pを求める式に書き換えると (A−22) 式が得られる。

$$P=\frac{(g^2-1)R+\sqrt{(g_0^2g^2-1)(g^2-1)}}{g^2-1+\sqrt{(g_0^2g^2-1)(g^2-1)}}$$

$$\therefore \boxed{\cos\omega_0 T=\frac{(g^2-1)\cos\omega_1 T+\sqrt{(g_0^2g^2-1)(g^2-1)}}{g^2-1+\sqrt{(g_0^2g^2-1)(g^2-1)}}}$$