Vol. 9 No. 43　　電波研究所季報　　July 1963
pp. 155—165

研　究

UDC 534.78

“合成による分析”法によるホルマント周波数の抽出

角川靖夫*　　中田和男*

FORMANT FREQUENCY EXTRACTION BY “ANALYSIS-BY-SYNTHESIS” TECHNIQUE

By
Yasuo KADOKAWA and Kazuo NAKATA

The importance and usefulness of analysis-by-synthesis technique in the analysis and recognition of speech have been recognized. The problems on the application of the method to formant frequency extraction were re-examined relating to the characters of our analyzer.

It is found that the additional process of pitch frequency adjustment is necessay in order to improve the accuracy of extraction and an efficient method of this pitch frequency adjustment is developed. This improvement can make acceptable the process not only in the case of male voices but in the case of female voices without any deterioration of results.

The over-all estimation of accuracy achieved by the method is better than that of the moment calculation method and can approach to the same order of D. L. of formant frequency by perception.

The whole process of automatic extraction of formant freqencies by analysis-by-synthesis technique was programed and tested with the practical input speech spectra.

The time scale of the process by NEAC-2203 is of the order of 10^5 and unrealistic in a sense even with making a full use of the first good approximatons obtained by the moment calculation method.

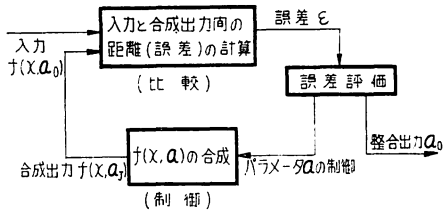
The extraction of formant frequencies on the full scale is scheduled to be carried on by the new computer NEAC-2206 in the future.

* 情報処理研究室

1. 緒　言

"合成による分析" 法 (Analysis-by-Synthesis Techniques) というのは，M.I.T. の K.N. Stevens 教授らによって彼らの音声の分析および識別の研究における一貫した情報処理の手法として，最近発表された方法であるが[1][2]，これはただ単に音声の分析，識別の研究に有効な手段であるというだけでなくて，パターン認識一般，さらにはそのほかの分野にも多くの有効な適用範囲がある方法と考えられる。

その原理と適用条件を一般的な形でのべると次のようにいうことができよう。パラメータ群 $\boldsymbol{a}_I\{a_{i1}, a_{i2}, \cdots, a_in\}$ をもつ関数の集合 $\{(x, \boldsymbol{a}_I)\}$ があって，そのうちの一つ $f(x, \boldsymbol{a}_0)$ の値が $f(x)$ の定義域で与えられ（観測され）たとき，それを規定している特定のパラメータの実現値 $\boldsymbol{a}_0$ を求めたい。このような場合次のような条件が満たされていれば，入力 $f(x, \boldsymbol{a}_0)$ と合成出力 $f(x, \boldsymbol{a}_I)$ の間に適当な距離測度をきめることによって第1図のような active analysis (or Analysis-by-Synthesis) process によって $\boldsymbol{a}_0$ を求めることが可能であり，また有効な方法でもある。*



第1図　合成による分析法の原理

すなわち，(1) $f(x, \boldsymbol{a})$ の関数形が既知である。(2) $\boldsymbol{a}_I$: $\{a_{i1}, a_{i2}, \cdots\cdots, a_{in}\}$ の要素の数 n が比較的大きく，その関係が複雑で，簡単に分離することができない。(3) $\boldsymbol{a}$ がほぼ連続的に変化する。

この方法の特色は，分析すべき対象 $f(x, \boldsymbol{a}_0)$ の operational な発生モデルが分析過程内に内そうされており，入力と合成出力の feedback 的な比較と制御によって $\boldsymbol{a}_0$ が漸進的に抽出されるという点にある。

音声の分析，識別においてこのような手法が提案されるに至った背景には，次の三つの重要な基礎的発展があったものと考えられる。

(1) 音声発生理論の形成，発展：G. Fant らによって形成された音声の発生理論が最近にいたってほぼ確立され[3]，周波数スペクトルの次元で operational な音声発生のモデルを提供することができるようになった。

(2) Articulatory Reference Theory of Speech Recognition の提唱：人間の音声認識過程のモデルとして発声器官の発声運動 (Articulation) を媒介として認識が行なわれるとする説が，従来の合成音声の聞き取り結果を基礎として唱えられ[5]，人間を一つの到達目標とする言語オートマトンの研究において，合成による分析法に原理的なモデルと思想的な支持を与えた。

(3) 計算機およびその入出力装置の発達：これによって合成による分析法を現実的な time-scale で実現するための技術的な手段が与えられた[5][6]。

以上の考察からもわかるように，この合成による分析法は音声の分析識別の研究において，特にその計算機による情報処理において，重要かつ有効な手法と考えられたので，計算機による音声の分析，識別の研究の一つとしてわれわれもこの方法をとり入れ，その第一歩として，音声のスペクトルからそれを規定するホルマント周波数を抽出する過程（母音型有声音）について検討し，自動抽出プログラムによる実測を行なったのでここに報告する。

特に M.I.T. との比較においてわれわれが検討し，改良したと思うのは次の諸点である。

(1) ホルマント周波数の収斂過程の第1近似値（出発点）としてモーメント計算法によるホルマント周波数の抽出値を用い，収斂過程の短縮をはかったこと。

(2) スペクトル分析濾波器の特性と関連して，音源基本周波数整合過程を導入し，ホルマント周波数の抽出誤差を軽減したこと。

(3) ホルマント周波数抽出過程を自動化し，さらに分析すべき母音型有声音区間を自動的に選定するようにしたこと。

結論的にいって，±5cps 精度の音源基本周波数の整合を含めて，第1ホルマント周波数で ±10cps以内，第2ホルマント周波数で ±20cps以内，第3ホルマント周波数で ±20cps 程度の精度で各ホルマント周波数を自動的に抽出することができ，この方法の有効性を確認した。

2. 合成による分析法による音声のホルマント周波数抽出の問題点 [2][7]

2.1. 整合させるべきパラメータ

音声発生理論によれば，定常的な母音型音声の周波数スペクトル $V(f)$ は(1)式のようにあらわされる。

---

* このほかに，パラメータ間にその変化による $f(x, \boldsymbol{a})$ の変化率すなわち $\partial f/\partial a_i$ に大小の順序があることが望ましい。

$$V(f)=S(f, f_0, \alpha)\cdot T(f, \{F_i, B_i\})\times R(f, \beta) \quad (1)$$

ここで $S(f)$ は音源のスペクトル，$T(f)$ は声道の伝達関数のスペクトル(共振特性)，$R(f)$ は唇からの音波の輻射特性（f は周波数)。

(1)式を規定するパラメータ，すなわち合成による分析法において整合させるべきパラメータは次のとおりである。

(i) $\{F_i, B_i\}$, $(i=1, \sim 3$ or $4)$ ：第1から第3（または第4）までのホルマント周波数とその共振帯域幅。

(ii) f_0：有声音源の基本周波数（ピッチ周波数)。

(iii) α ：音源スペクトルの特異性をあらわすパラメータ

(iv) β ：高次ホルマントの効果の補償特性をあらわすパラメータ

普通は(1)式のような音声スペクトルを測定するのに，帯域濾波器群が用いられるが，その場合には実測データとよい一致をとるために，(1)式の上にさらに各帯域濾波器の特性を考慮した Filter Simulation をほどこす必要がある。

M. I. T. での合成による分析法によるホルマント周波数の抽出過程では，暗黙のうちに次のことが仮定されている。

(1) 上述の各パラメータのうちで最も重要な（変化の効果：$\partial f/\partial a_i$ の大きい）ものは $\{F_i\}$ であって，α, β は特定の個人についてはほぼ一定であり，またその個人差の効果は二次的であり，f_0 $\{B_i\}$ とともにその平均値的な第1近似値から $\{F_i\}$ の整合をとり始めて unique に $\{F_i\}$ の真の値に漸近することができる。

(2) 1次パラメータ $\{F_i\}$ と2次パラメータ $\{B_i\}$, α, β との整合過程をくり返すことによって，最終的には unique に特定の $\{F_i, B_i\}$, α, β の値に収斂しそれが求める値である。

2·2. 分析装置の性能の差異による問題点

われわれの利用しうる音声分析用計算機入力装置[6]の性能と M.I.T. のもの[2]との間には第1表に示すような差異がある。この性能の差異による問題点のうちで特に問題となると思われるのは次の点である。

(1) 基本周波数（ピッチ周波数）未整合による誤差

分析帯域濾波器群の特性との関係で，われわれの場合音源基本周波数 f_0 を未整合のままで，はたして $\{F_i\}$ の整合過程が unique に $\{F_i\}$ の真の値に漸近するか，またその誤差はどの程度か。

(2) 入力レベル数の不足による誤差

8ビット程度の入力レベル数でどの程度の抽出精度がえられるか。

(3) 抽出誤差軽減のため音源基本周波数 f_0 を整合させる必要があるとなった場合にどのようにして能率的に f_0 を整合させるか。

2·3. 一般的に検討すべき問題点

上述のほかに一般的に検討すべき問題点として考えたのは次の諸点である。

(4) 整合させるべきパラメータの特性に応じてその整合度を判定する最適な measure はどうか。

(5) ホルマント周波数整合過程を最少ステップ数で収斂させるための最適な strategy はどうか。

(6) 総合的なホルマント周波数抽出精度およびその所要時間はどのくらいか。

以上の諸点について検討した結果を次にのべる。*

第1表 音声分析用計算機入力装置の性能の比較

項目＼研究所	M. I. T.	電波研究所 (R.R.L.)
分析チャネル数 （母音分析の場合）	36 (150cps～7025cps) 24 (150cps～3000cps)	26 (200cps～5900cps) 23 (200cps～3900cps)
分析用帯域濾波器幅	1600cps まで 100cps （それ以上対数的に漸増）	1200cps まで 100cps 2700cps まで 150cps それ以上 400cps
分析用帯域濾波器特性	single—tune filter	sharp cut-off filter （隣接チャネル中心で−20dB）
A-D変換レベル数	11ビット（符号を含む）	8ビット（符号を含む）
サンプリング周期	8.6 ms（同時サンプル）	10 ms (hold なし)
計算機（主メモリー）	TX-O（加算 12μs） コアー（8000語）	NEAC-2203（加算 900μs） ドラム（2000語）

3. 問題点の検討結果

ホルマント周波数の整合を行なう過程の整合度の measure としては，入力スペクトルパターン $\{A_j\}$ と合成出力スペクトルパターン $\{S_j\}$

---

* Analysis-by-Synthesis によるホルマント周波数抽出の一般的な内容については，文献（2），(7) を参照されたい。

の平均レベル整合のもとでの自乗誤差をとった。すなわち $\sum_j A_j=\sum_j S_j=0$ の条件のもとで $\sum_j(A_j-S_j)^2=\mathcal{E}^2$ を最小にする。このことは結局 $\{A_j\}$ と $\{S_j\}$ との間の相互相関係数 r を最大にすることに等しくなるので，実際にはそれを criterion として用いた。すなわち

$$r=\frac{\sum_j A_jS_j}{(\sum A_j^2\cdot\sum S_j^2)^{\frac{1}{2}}}$$

を整合の measure とした。

第 2 表　音源基本周波数未整合でのホルマント周波数整合の度合の 1 例
（相互相関係数 r で表示）

(a)　粗い整合

F_{10}	F_{20}	F_1			F_{10}	F_{20}	F_2		
		$F_{10}-100$	F_{10}	$F_{10}+100$			$F_{20}-200$	F_{20}	$F_{20}+200$
200	1500		0.7371	0.9041	500	800		0.8883	0.7915
300		0.8004	0.9977	0.6409		1000	0.8380	0.8083	0.7843
400	⌒	0.5370	0.8719	0.7080	⌒	1200	0.7199	0.8432	0.7012
500	固	0.7053	0.8158	0.4670	固	1400	0.7809	0.8423	0.7311
600		0.6155	0.9915	0.6205		1600	0.7052	0.8229	0.7179
700		0.4618	0.8108	0.7482		1800	0.4887	0.8106	0.5302
800	定	0.6739	0.8233	0.6148	定	2000	0.5761	0.8492	0.7222
900	⌒	0.6688	0.9898		⌒	2200	0.5985	0.9233	0.7793
						2400	0.7305	0.9753	

入力のホルマント周波数，F_{10}, F_{20}，　音源基本周波数，　f_0=150 cps

合成のホルマント周波数　左表　F_1, F_{20}（1500 cps 固定）　右表　F_{10}（500 cps 固定）　F_2　f_0=100 cps

3.1.　音源基本周波数未整合による誤差

分析帯域濾波器群の特性とに関連して，われわれの場合一番問題になるのが音源基本周波数 f_0 未整合の効果である。この f_0 未整合による誤差を検討するのに次のような方法をとった。

付録に示すような音声スペクトル合成式で合成した音声スペクトルを入力として用い，ただ f_0 とホルマント周波数 $\{F_i\}$ のみを変化させ，*その間の整合度を(2)式の相関係数 r によって計ってみた。その 1 例を第 2 表に示す。

このようなシミュレーションによる検討の結果，2 次パラメータとしては f_0 のみが未整合の状態で，第 1 ホルマント周波数で ±25cps 程度，第 2 ホルマント周波数で ± 50cps 程度の範囲内には unique に接近しうることがわかった。さらにその範囲内で f_0 未整合でどの程度の誤差を抽出結果に生じるかを前と同様なシミュレーションによって検討した。その実例の一つを第 2 図(次頁)に示す。

これらの結果から入力の f_0 が 100～140cps の範囲では，合成の f_0 を 100cps に固定して f_0 未整合のままでも，第 1 ホルマント周波数で ±10cps 以内，第 2 ホルマント周波数で ±20cps 以内の抽出精度がえられるが，入力の f_0 が 160 cps 以上になると f_0 未整合（合成側の f_0 = 100 cps）のままでは誤差が多くなることがわかった。

(b)　比較的細かい整合

F_2	F_1=840	F_1=880	F_1	F_2=1100	F_2=1300
1100	0.8624	0.8760	840	0.8624	0.8740
1150	0.9185	0.9248	850	0.8758	0.9150
1200	0.9635	0.9829	860	0.8819	0.9307
1250	0.9616	0.9616	970	0.8819	0.9386
1300	0.9254	0.9254	880	0.8760	0.9470

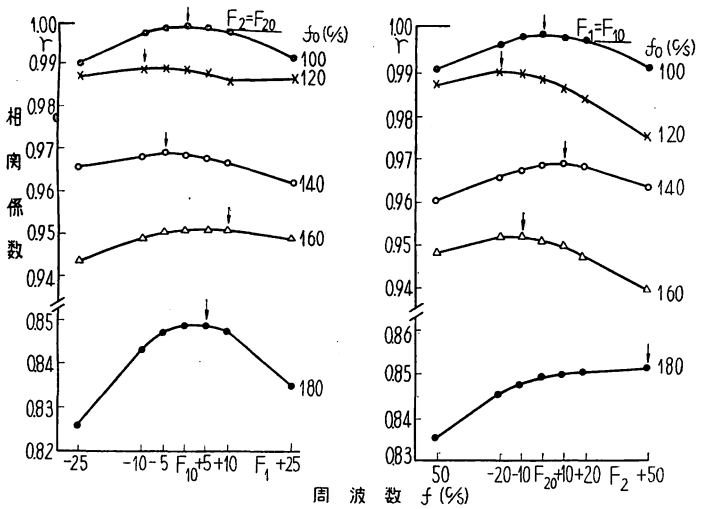
入力のホルマント周波数，F_{10}=860cps, F_{20}=1200cps, f_0=150cps

合成のホルマント周波数　左表　F_1=840 cps と 880 cps, F_2　右表　F_1, F_2=1100 cps と 1300 cps　f_0=100 cps

3.2.　入力レベル数の不足による誤差

入力レベルの変化による誤差を検討するために次のような方法をとった。3.1 の場合と同様計算機内部で合成したスペクトルを入力として用いるが，その入力の最大値のレベルを100（7 ビットは最大 127 に相当），50，25，10 の 4 段階に変化させて f_0 未整合の状態でホルマント周波数の整合を行なった。その実例の一つを第 3 図（次頁）に示す。これらの結果から k=10（入力の最大振幅 10）の場合を除いては，入力レベルの変化より f_0 未整合の効果の方が大きいことがわかった。また最大振幅レベルの変化による整合状態の著しい変化はみとめられないこともわかった。

＊　他のパラメータは整合の状態にある。



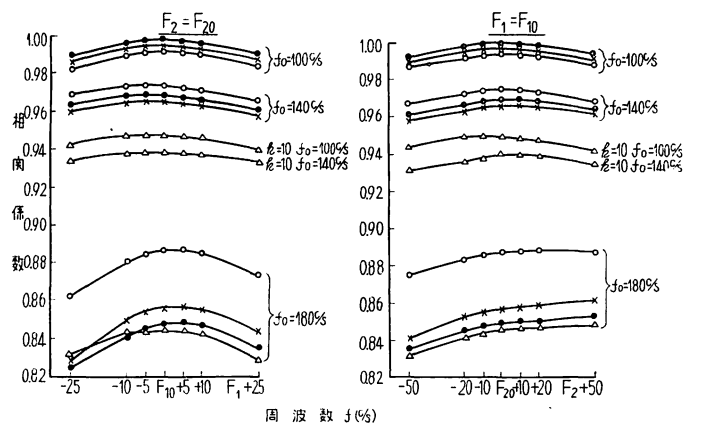
第2図 音源基本周波数 f_0 の未整合でのホルマント周波数抽出精度の1例

最大振幅入力レベル k=100 一定，合成の f_0=100c/s 一定，$F_3=F_{30}$=2500c/s，$F_4=F_{40}$=3500c/s，$F_5=F_{50}$=4500c/s

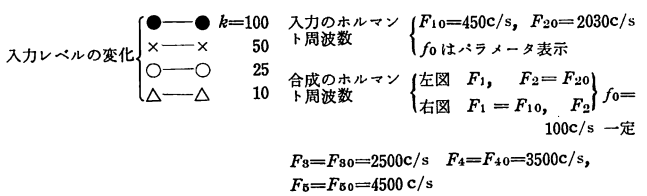
入力のホルマント周波数 F_{10}=450c/s，F_{20}=2030c/s，f_0はパラメータ表示

合成のホルマント周波数 左図 $F_1, F_2=F_{20}$ (2030 c/s)，右図 $F_1=F_{10}$ (450c/s)，F_2

↓印は最大相関すなわち最適整合の値を示す。



第3図 最大振幅入力レベルの変化によるホルマント周波数抽出精度の1例



3.3. 音源基本周波数 f_0 の整合

3.1 の検討によりわれわれの場合には分析帯域濾波器群の特性と関連して入力音声スペクトルの音源基本周波数 f_0 が 150cps 程度以上のときは f_0 の整合をも行なわなければ十分な精度で所望のホルマント周波数を抽出することができないことがわかったので，次に f_0 の能率的な整合過程を検討した。種々検討した結果，f_0 未整合の程度を示す measure としてはスペクトルの極大値 (local peaks) の数，および f_0 = 100cps の合成出力* との間の変化分(3)式で定義されるような D') が有効であり，f_0 整合には D' を最小にするという criterion が有用であることがわかった。

$$D'=\sum_j|(A_j-S_j)-(A_{j+1}-S_{j+1})|$$
$$=\sum_j|A_j-A_{j+1})-(S_j-S_{j+1})| \quad (3)$$

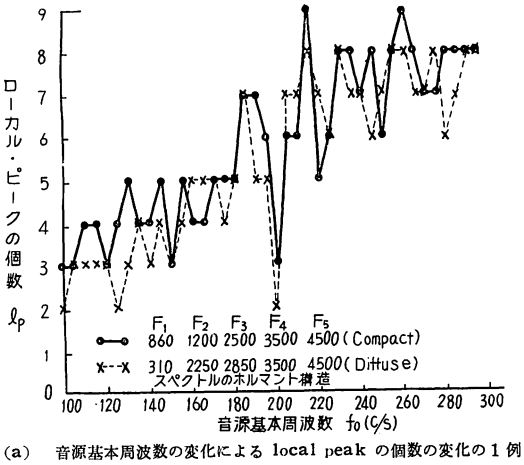
シミュレーションによる検討結果の1例として第4図(a)，(b)(次頁) に入力スペクトルの f_0 の変化による local peaks l_p の個数と D' の値の変化例を示す。これらの実例からこの二つの量がその音声のホルマント構造にはほぼ無関係で f_0 の変化のみによって変化する量であり，したがって f_0 の未整合を検出するのによい measure であることがわかる。

これらの結果から入力スペクトルの local peaks の個数 l_p と f_0=100 cps の合成出力との間の D' の値から，入力の f_0 の値を $f_0<160$ cps，160 cps$\leq f_0\leq$200 cps，$f_0>200$ cps の3段階程度にわけて推定することができる。

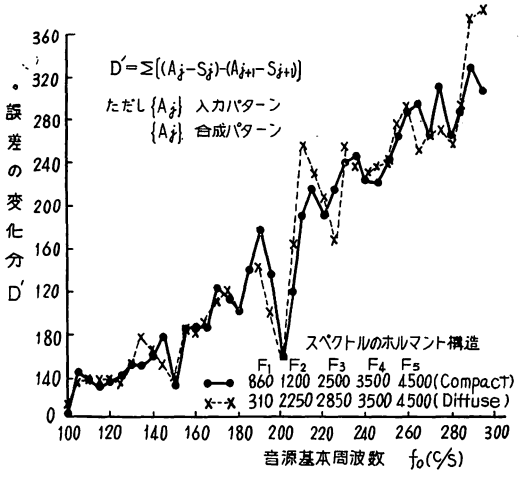
$f_0<160$ cps のときは f_0 を整合する必要はなく，$f_0>200$ cps のときは入力スペクトルの local peaks の位置（その周波数）から f_0 の存在範囲を決めることができるから(8)，その範囲中で D' をmeasure として f_0 の整合を行なえばよい。この中間では f_0 の存在範囲を決めることはできないので，一応 150cps と 200cps の間で5 cps ステップで D' を measure として入力の f_0 を探すことにした。

3.4. 整合させるべきパラメータの特性に応じた最適の整合 measure の選定

＊ ホルマント周波数は少なくとも近似的に整合している必要がある。



(a) 音源基本周波数の変化による local peak の個数の変化の1例



(b) 音源基本周波数の変化による誤差の変化分 D' の変化の1例

第4図 音源基本周波数の未整合検出の measures

入力スペクトルの音源基本周波数 f_0 (変化), 合成スペクトルの音源基本周波数 f_0=100c/s (固定)

すでに述べたように，ホルマント周波数の整合には(2)式の相関係数 r，音源基本周波数の整合には(3)式の D' を用いることにした。

3.5. 最小ステップ数で収斂させるための strategy

各パラメータの整合過程を最小ステップ数で誤差少なく収斂させるためには次の諸点を考慮する必要がある。

(1) 第1近似値（整合過程の出発点）としてできるだけ真の値に近い値を利用すること。

(2) パラメータ間の整合順序をその影響力の大きいものから整合させてゆくこと。

(3) 整合探索の歩幅を漸減していくこと。

最も重要かつ影響力の大きいパラメータは当然ホルマント周波数 $\{F_i\}$ である。そこで $\{F_i\}$ の第1近似値としてはわれわれの考案による“モーメント計算によるホルマント周波数抽出法”[8]による値を用いることにした。この方法による抽出ホルマント周波数の誤差は $f_0 \leqq 200$ cps で第1ホルマント周波数については±5％，第2ホルマント周波数については±10％以内であり，$f_0 > 200$ cps のときでも前者で±10％，後者で±20％程度であり，第1近似値としては十分の精度をもつ値である。* 第3ホルマント周波数についてはこのように利用しうる適当な近似値がないので 2500cps ～3000cps の間で探すことにした。

パラメータ間の整合順序としては，ホルマント周波数が第1であるが，上にのべたような精度のよい第1近似値を利用することができるので，まず音源基本周波数 f_0 の整合を行なうこととした（第5図参照）。

またホルマント周波数の間でも，最も相関係数 r に大きく影響するものから整合させるのが能率がよい。一般的には第1ホルマントの周波数が最も影響力が大きく，第1→第2→第3の順が有効だが，第1ホルマントの周波数が低く，第2ホルマントの周波数が高い（いわゆる diffuse な）場合には第2ホルマントの影響の方が強い[9]。そこで簡単なルールとしてモーメント法による第1近似値の第2ホルマントの周波数が2000cps より高いときには第2→第1→第3の順に整合をとることにした。

また音源基本周波数 f_0 の整合についても3.3にのべたような方法により能率的に整合させることを考えた。

4. ホルマント周波数自動抽出過程およびその実測例

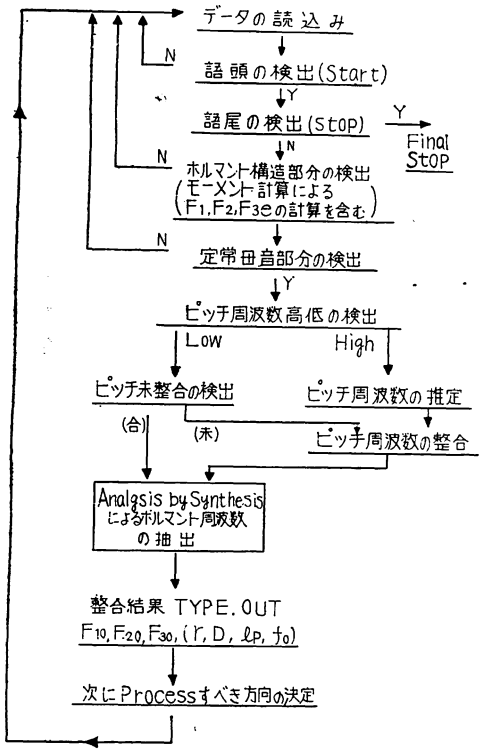
4.1. ホルマント周波数自動抽出プログラム

以上のべたような諸問題点の検討結果およびすでに別に行なった音声スペクトルの分析，音韻分類プログラムの結果から，総合的な母音型有声音のホルマント周波数の自動抽出過程をプログラムした。その概要を第5図に，ホルマント周波数抽出部分の概要を第6図に示す。

また音声スペクトル合成のための主要な計算式，filter Simulation の計算法などを付録に示す。

モーメント計算によるホルマント周波数の抽出法およ

* 計算所要時間は数秒である。



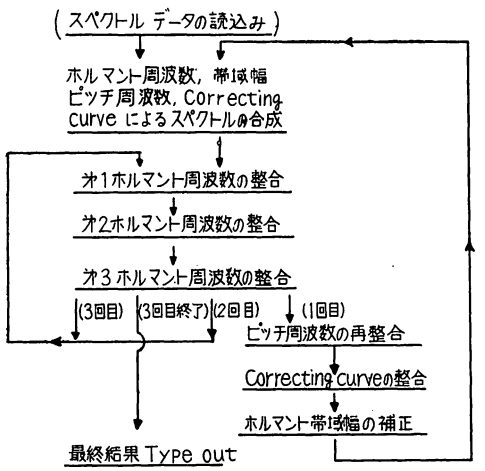
第5図　合成による分析法によるホルマント周波数抽出プログラム

注：(1) F_1, F_2, F_{3e} は モーメント法によるホルマント周波数の第1近似値

(2) F_{10}, F_{20}, F_{30} は合成による分析法による整合ホルマント周波数

(3) f_0 は 音源基本周波数整合を行なった場合の整合音源基本周波数

(4) r は整合時の相関係数，l_p は local peak の個数，D' は整合時の誤差の変化分



第6図　合成による分析法によるホルマント周波数抽出過程

注 1：各段階におけるホルマント周波数整合の 精度と最小探索範囲を右側の表に示す。

び定常母音部分の検出法などについては文献 (6), (8), (10) を参照されたい。

4.2. 実測例

この合成による分析法によるホルマント周波数自動抽出の実例をスペクトルの形で第7図に，5母音についての1例を第3表に，また単音節全体についての実例を第8図に示す。なおこれらの実測例ではホルマントの帯域幅 $\{B_i\}$，第4，第5ホルマント周波数の整合は行なっていない。

総合的なホルマント周波数精度は3章のシミュレーションによる検討結果および実測値のソナグラム，さらにモーメント法によるホルマント周波数の抽出結果との比較などから，第1ホルマント周波数で ±10cps 以内，第2ホルマント周波数で ±20cps 以内と考えられ，Difference Limen による人間の弁別閾値3～5％[11]とほぼ同程度と考えられる。

所要時間は NEAC-2203 で1回のスペクトル合成に約30秒を要するため，1回の入力データの処理（第1～第3のホルマント周波数とそのときの最大相関値 r および整合を行なった場合には，その音源基本周波数 f_0 とそのときの最小 D' の値をタイプアウトするまで）に約20分を必要とする。

計算所要時間を短縮する方法として M.I.T. では次のような方法を用いた[2][7]。

すなわち適当な帯域幅をもつホルマント周波数を0～4000cps の間に78組選び，その包絡線を計算して，それに Filter Simulation をほどこしたものを store しておく。そしてスペクトルの合成には #1～#78 の曲線の

ホルマント周波数抽出の精度と最小探索範囲

回数	第1ホルマント F_1 (c/s)		第2ホルマント F_2 (c/s)		第3ホルマント F_3 (c/s)	
	精度	範囲	精度	範囲	精度	範囲
第1回目	±10	±20	±20	±10	±50	2500～3000
第2回目	± 5	± 5	±10	±10	±20	±40
第3回目	± 5	± 5	±10	±10	±20	±20

注 2：　最小探索範囲(両端を除く)のうちに r の最大値がないときは，r の増加方向に同一の精度で最大値がみつかるまで探索をすすめる。

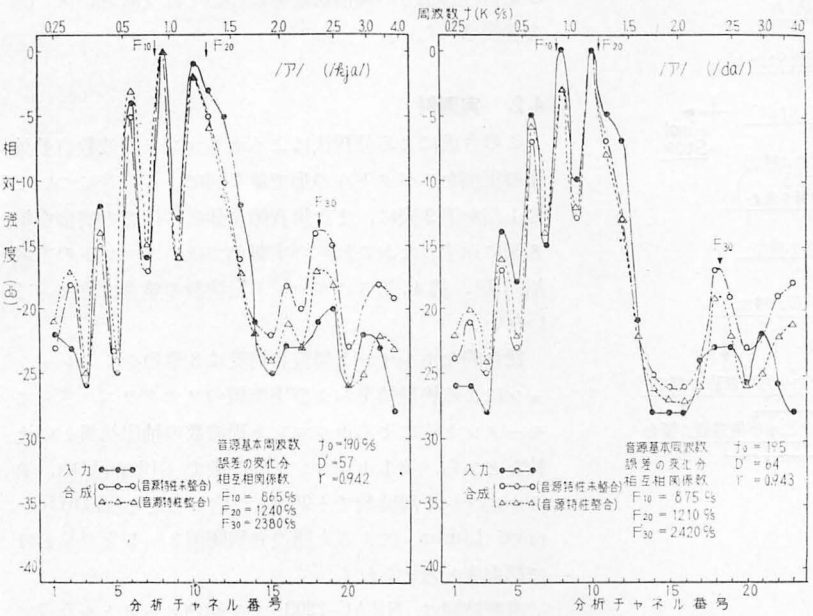
注 3：　ホルマント周波数整合の測度は入力パターン $\{A_j\}$ と合成パターン $\{S_j\}$ との相互相関係数 r による（r の最大値が整合条件）

$$r=\sum_j A_jS_j/\left[\left(\sum_j A_j^2\right)\left(\sum_j S_j^2\right)\right]^{\frac{1}{2}}$$

ただし　$\sum_j A_j=\sum_j S_j=0$

注 4：　音源基本周波数整合の測度は $\{A_j\}$ と $\{S_j\}$ との間の誤差の変化分

$$D'=\sum_j|(A_j-A_{j+1})-(S_j-S_{j+1})|$$



第7図　整合の実測例（スペクトル表示）

↗さらに大となり，現在のわれわれの計算機 NEAC-2203 では実行不可能である。

そこで次のような2通りの計算法を試みた。

(1) 正確な計算法：音源基本周波数の決定→スペクトル・エンベロープの計算（ピッチの各高調波において）→フィルターシミュレーション（ピッチの各高調波において）→合成出力。

(2) 近似計算法：音源基本周波数の決定→フィルターシミュレーション（ピッチの各高調波において）→スペクトルエンベロープの計算（各分析濾波器の中心周波数において；23個）→合成出力。

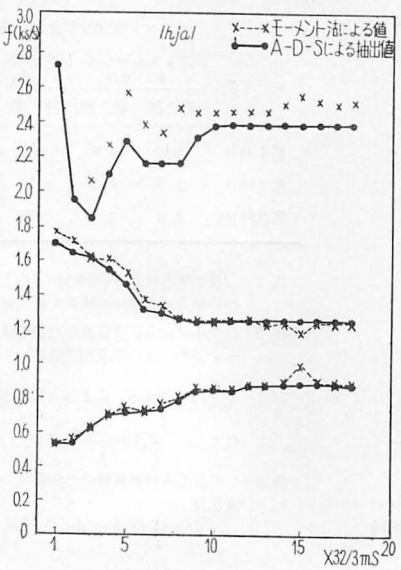
この二つの計算法の所要時間

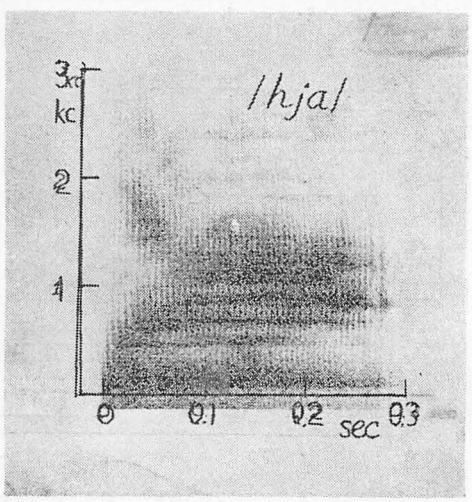
第3表　5母音の実測例

入力母音	モーメント法による第1近似値			合成による分析法による結果										
				音源基本周波数未整合（f_0=100c/s）					音源基本周波数整合					
	F_1	F_2	F_3	F_1'	F_2'	F_3'	r	D'	f_0	F_{10}	F_{20}	F_{30}	r	D'
ア（ダ）	860	1260	(2500)	850	1230	2350	0.856	133	195	875	1210	2420	0.921	65
イ（イ）	310	2250	(〃)	310	2030	2430	0.736	162	195	270	2030	2900	0.739	159
ウ（プ）	390	1170	(〃)	385	1040	2540	0.790	178	160	390	1080	2480	0.763	131
エ（デ）	530	1170	(〃)	520	1800	2240	0.653	194	190	530	1780	2200	0.854	131
オ（コ）	470	770	(〃)	455	700	2880	0.793	179	195	465	730	2840	0.800	179

中から第1～第4のホルマントに相当する4個を選んで加え合せる（dB 変換されているから）。

この方法では，計算時間は非常に速くなるが，必要な記憶容量が大となる（23チャンネルまで計算するとして，23 × 78 = 1794 語が必要）。音源基本周波数を可変にしようとすると必要な記憶容量は↗





第8図　単音節の実測例→

第 4 表　異った計算法による所要時間と精度の比較

(a)　計算所要時間の比較（1回のスペクトル合成）

音源基本周波数 f_0 (c/s) ＼ 計算法	Ⅰ (s)				Ⅱ (s)				時間比率 (%)
	S	F	r	T_1	S	F	r	T_2	T_2/T_1
100	5.5	16.5	11.0	33.0	2.5	6.5	11.0	20.0	61
150	3.5	11.5	11.0	26.0	2.5	5.5	11.0	19.0	73
200	2.5	9.0	11.0	22.5	2.5	4.5	11.0	18.5	82

r：　dB 変換　5.5sec，相関の計算 5.5 sec
S：　1回のスペクトル合成式の計算
F：　フィルターシミュレーション
T_1, T_2：　Ⅰ, Ⅱの方法による1回のスペクトル合成の全時間（sec）

(b)　ホルマント周波数（c/s）の誤差の1例
（合成入力による）

入力のホルマント周波数 F_n (c/s)	計算法	入力の音源基本周波数 f_0 (c/s) 100	120	140
$F_{10}=450$	Ⅰ	0	−10	− 5
	Ⅱ	+30	+30	−20
$F_{20}=2030$	Ⅰ	0	−20	+10
	Ⅱ	+30	+20	+40
$F_{30}=2500$	Ⅰ	—	—	—
	Ⅱ	+20	+20	+20

(c)　実測の1例

A—b—S による値 ＼ 入力母音 ＼ 計算法		音源基本周波数 f_0 (c/s)	ホルマント周波数 F_n (c/s) F_{10}	F_{20}	F_{30}
ア（ダ）	Ⅰ	795	875	1210	2420
	Ⅱ	195	905	1220	2400
ア（ヒア）	Ⅰ	190	865	1240	2380
	Ⅱ	190	895	1220	2380

注：　Ⅰ　正確な計算法，　Ⅱ　近似計算法

の比較の1例を第4表(a)に示す。この表からもわかるように，音源基本周波数（ピッチ周波数）が高くなると考慮する帯域（90cps～4340cps）内の高調波数が少なくなり，近似計算法による所要時間の軽減効果が減少する。

またこの二つの方法によるホルマント周波数抽出誤差の1例を第4表(b)に，実測の入力データに対する抽出値の比較の1例を第4表(c)に示す。

結論的にいって，分析濾波器の特性が sharp cut-off であるわれわれの場合には，上述の近似計算による所要時間の軽減率は少なく，誤差の増加が著しいので，音源基本周波数が低い場合（f_0=100cps）を除いては有効とはいえない。

5. 結　　言

音声の分析，識別における“合成による分析”法の重要性と有用性をみとめ，われわれの利用しうる分析装置との関連においてその問題点を検討し，その検討結果にもとづいて音声スペクトルからそのホルマント周波数を自動的に抽出するプログラムをつくって実測を行なった。

われわれの場合，十分な抽出精度をうるためには，分析帯域濾波器群の特性に関連して音源基本周波数の整合をも行なう必要のあることがわかり，そのための過程を加えた。このためわれわれのプログラムは原理的には男声にも女声にも適用することができるように一般化された。

総合的な抽出精度としては，第1近似値として用いたモーメント法による抽出精度を一段と高め，Difference Limen による人間の弁別閾値とほぼ同程度の精度まで得られたものと考えられる。ただ計算所要時間が NEAC－2203 程度の計算機（ドラムメモリー）では第1近似値としてモーメント法によるホルマント周波数の抽出値のようなよい近似値を用いても，time scale で 10^5 order となり，やや非現実的といわなければならない。またプログラム語数も相当に多くなり NEAC－2203 程度（2000語）では第6図に示す部分だけでほぼ一ぱいとなる。われわれの計画としては新たに設置予定のＮＥＡＣ－2206（コアメモリー，4000語，磁気テープつき）によって本格的な実測および今後の研究を行なう予定である。

今後の問題としては，さらに分析の過程を一段すすめて，発声時の声道の形およびその運動（articulation）のレベルで，音声の分析および識別の研究[2]を行なうように準備をすすめている。

最後に常に御指導をいただいている上田所長，河野次長，尾方室長に感謝し，また共同で研究をすすめている鈴木技官の協力に謝意を表します。また M. I. T. での計算の詳細な内容について教えていただき，われわれの方法についても種々討論していただいた東大工学部藤崎博也氏に厚く御礼申し上げる。

付　　録

1. 音声スペクトルの合成計算式

G. Fant の音声発生理論によれば[3]，空間に音波として発生された準定常的な音波波形の周波数スペクトルは（a—1）式のようにあらわされる。

$$V(f)=S(f)\cdot T(f)\cdot R(f), \qquad \text{(a—1)}$$

ここで $S(f)$ は音源のスペクトル，$T(f)$ は声道の伝達関数のスペクトル，$R(f)$ は唇からの音波の輻射特性をあらわす。

有声音の場合には，$S(f)$ はその音源基本周波数 f_0 の高調波からなる線スペクトル構造をもち，そのスペクトルエンベロープは平均 −12dB/oct の勾配で周波数とともに低下する。また $R(f)$ は 6dB/oct の高域強調と考えられる。われわれの計算では $S(f)$ と $R(f)$ とを一括して Fant による半実験的な近似式 (a—2) によった。

$$S(f)\cdot R(f)=\frac{f/100}{1+(f/100)^2} \qquad \text{(a—2)}$$

母音型音声の場合には $T(f)$ は一般に極（共振）のみを有し，(a—3) 式のようにあらわせる。

$$T(f)=\prod_{s=1}^{\infty}\frac{F_i^2+(B_i/2)^2}{\sqrt{(f+F_i)^2+(B_i/2)^2}\cdot\sqrt{(f-F_i)^2+(B_i/2)^2}} \qquad \text{(a—3)}$$

ここで F_i は i 番目の極（ホルマント）の周波数，B_i はその帯域幅に相当する。母音型有声音声の場合，その分析，識別に重要な周波数範囲は 3000 cps 程度までであり，その範囲では $i=1\sim3$ or 4 すなわち第1～第3もしくは第4ホルマントまでを規定すれば十分であり，高次のホルマントの効果は一括して補償項として取扱うことができる。われわれの計算では第5ホルマントまで (a—3) 式の型で考慮し（ただし $F_4=3500$cps，$F_5=4500$cps に固定），高次ホルマントの効果は (a—4) 式のような補償項 $K_5(x)$ として加えた。

$$K_5(x)=\exp\left[\left\{\frac{\pi^2}{8}-\sum_{n=1}^{5}\frac{1}{(2n-1)^2}\right\}x_1^2+\frac{1}{2}\left\{\frac{\pi^4}{96}-\sum_{n=1}^{5}\frac{1}{(2n-1)^4}\right\}x_1^4\right] \qquad \text{(a—4)}$$

(13)

ここで $x=f/F_1\simeq 4fl_e/c$

l_e は発声時の等価的な声道の長さで，平均的な計算では neutral な母音（$F_0=500$cps or $l_e=17.6$cm）で代表させている。

ホルマント帯域幅 B_i の第1近似値としては，$B_1=50$ cps，$B_2=80$ cps，$B_3=100$ cps，$B_4=B_5=200$cps を用いている(14)。

また $S(f)$ の特異性を示す式としては (a—2)式を次のように修飾して用いている。

$$S'(f)\cdot R(f)=\frac{f/100}{1+(f/100)^2}\cdot\frac{1}{1+(f/f_c)} \qquad \text{(a—2)'}$$

ここで f_c を 40kc，4kc，2kc，1kc，0.5kc の5段階にわたって変えることによって5種類の correction curves とした。

実際には $S(f)$ の線スペクトル構造のため，これらの式はすべて音源基本周波数 f_0 の高調波周波数 $f=mf_0$ ($m=1, 2, 3, \cdots\cdots$) において計算することになる。

2. Filter Simulation の計算法

音声スペクトルは第1表にその主要な性能を示した分析用帯域濾波器群によって観測される。したがって入力データと合成出力のよい一致を求めるためには，付録1に示したような式によって計算された音声スペクトルに，さらにこの分析用帯域濾波器群を通った効果を加えなければならない。この計算を Filter Simulation といっており，原理的には次式の計算をすることになる。

$$S_i=\int_0^{\infty}\left[B_j(f)\cdot V(f)\right]^2 df \qquad \text{(a—5)}$$

ここで B_j は j 番目の帯域濾波器の通過特性である。$S(f)$ の高調波線スペクトル構造と $B_j(f)$ の特性との関係でやや手数のかかる計算になるが，その計算法のフローチャートを第9図に示す。

なお実際の整合過程はすべて dB 変換された電力スペクトルについて行なっている。

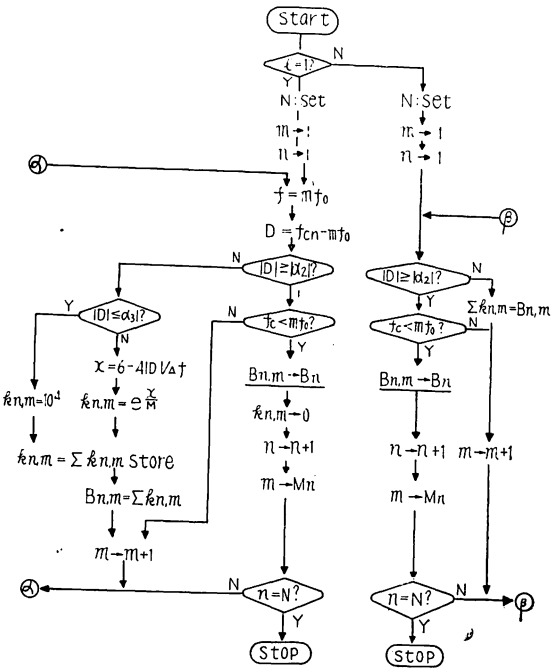
参照文献

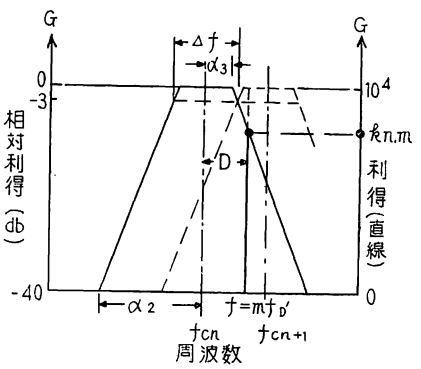
(1) K. N. Stevens; "Toward a Model for Speech Recognition", JASA., 32, 1, pp. 47-55, (Jan., 1960).

(2) C. G. Bell, H. Fujisaki et al; "Reduction of Speech Spectra by Analysis-by Synthesis Techniques", JASA, 33, 12, pp. 1725-1736, (Dec., 1962).

(3) G. Fant; "Acoustic Theory of Speech Production", Mouton & Co., 's-Gravenhage, (1960).

(4) A. M. Liberman, F. S. Cooper et al; "A Motor Theory of Speech Perception", Speech Communication Seminar, Stockholm, (1962).





f_0： 音源基本周波数
l： f_0が一定のときの filter Simulation の回数
m： 高調波の次数
n： 分析チャネルの番号
α_2： 通過帯域（−40dB）
α_3： 完全通過帯域（0dB）
Δf： 帯域幅（−3dB）
f_{cn}： 分析チャネルの中心周波数
N： 分析チャネルの個数（=23）
M_n： パラメター
$k_{n,m}$：利得係数
$\frac{1}{M}$=2.302585

第 9 図　Filter Simulation のフロー・チャート

(5) E. E. David, Jr.; "Digital Simulation in Research on Human Communication", IRE., 49. 1. pp. 319-329, (Jan., 1961).

(6) 鈴木, 角川, 中田, 前園; "計算機による音声の分析", 情報と制御の研究, 2, 3 号, pp. 27—40, (昭37).

(7) 藤崎博也; "電子計算機による母音のホルマント抽出", 情報と制御の研究, 2, 3 号, pp. 11—19, (昭37).

(8) 鈴木, 角川, 中田; "モーメント計算によるホルマント周波数の抽出", 音響学会誌, 19, 3, (昭38. 5 印刷中)

(9) 角川, 中田, 鈴木; "Analysis-by-Synthesis によるホルマント周波数の抽出", 38年電気四学会連合大会, No. 1159.

(10) 鈴木, 中田, 前園; "単音節の識別", インホメーション理論研究会資料, (昭38. 3).

(11) J. L. Flanagan; "Percepture Criteria in Speech Processing", Speech Communication Stockholm, (1962).

(12) J. M. Heinz; "An Analysis of Speech Spectra in term of a Model of Articulation", Speech Communication Seminar Stockholm, (1962).

(13) G. Fant; "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", Ericsson Technics, No. 1, (1959).

(14) H. K. Dunn; "Method of Measuring Vowel Formant Bandwidths", JASA, 33, 9, pp. 1737-1746, (Dec., 1961).