



⑫A **Terinzagelegging** ⑪ **8400552**

Nederland

⑲ NL

-
- ⑤4 **Systeem voor het analyseren van menselijke spraak.**
⑤1 Int.Cl.: G10L7/02.
⑦1 Aanvrager: N.V. Philips' Gloeilampenfabrieken te Eindhoven.
⑦4 Gem.: Ir. P.J.P.G. Simons.c.s.
Internationaal Octrooibureau B.V.
Prof. Holstlaan 6
5656 AA Eindhoven.

-
- ②1 Aanvraag Nr. 8400552.
②2 Ingediend 22 februari 1984.
③2 --
③3 --
③1 --
⑥2 --

-
- ④3 Ter inzage gelegd 16 september 1985.

De aan dit blad gehechte stukken zijn een afdruk van de oorspronkelijk ingediende beschrijving met conclusie(s) en eventuele tekening(en).

NV. Philips' Gloeilampenfabrieken te Eindhoven.

"Systeem voor het analyseren van menselijke spraak".

A. Achtergrond van de uitvinding

A.1 Gebied van de uitvinding

5 Systeem voor het analyseren van menselijke spraak voor het bepalen van de toonhoogte van spraaksegmenten onder toepassing van meer dan één toonhoogtedetectie-algoritme.

A.2 Beschrijving van de stand van de techniek

10 Een systeem zoals hierboven genoemd is bekend uit referentie D1. In het aldaar beschreven systeem wordt gebruik gemaakt van de autocorrelatiemethode, de cepstrummethode en van de laagdoorlaatgefilterde golfvorm. De keuze van deze methodes werd zoals in deze publicatie is aangegeven bepaald door de wens om redelijk onafhankelijke schattingen van de toonhoogte te verkrijgen.

15

De autocorrelatiemethode maakt direct gebruik van informatie uit het tijddomein (referentie D2), terwijl de cepstrummethode gebruik maakt van informatie uit het frequentiedomein. Andere methoden welke gebruik maken van informatie uit het frequentiedomein zijn bekend, bijvoorbeeld 20 de in referentie D3 beschreven harmonische zeefmethode. Daar wordt het amplitudespectrum bepaald voor een kort segment (40 ms) van het bemonsterd signaal en vervolgens wordt in het amplitudespectrum gezocht naar de frequentieposities van de significante pieken van de amplitude (significante piekposities) en tenslotte wordt -door wat wordt genoemd de 25 harmonische zeef- gezocht naar een toonhoogte waarvan de harmonischen het beste passen bij de significante piekposities van het amplitudespectrum.

Bij de hier genoemde methodes voor het bepalen van de toonhoogte in spraak doen zich voor elke methode karakteristieke problemen voor. Over 30 het algemeen kan men zeggen dat methodes opererend in de frequentiewereld nogal eens fouten maken voor hoge toonhoogtes en dat methodes werkend in de tijdwereld juist fouten maken voor lagere toonhoogtes en vaak veelvouden van de eigenlijke toonhoogte als toonhoogte aangeven.

8400552

B. Samenvatting van de uitvinding

De uitvinding beoogt een systeem van het in A.1 aangegeven type te verschaffen met twee detectie-algoritmes welke op optimale wijze complementaire toonhoogte-informaties verschaffen, welke gerekend over het gebied van lage tot hoge toonhoogten complementair zijn voor wat betreft de betrouwbaarheid van de informatie, waarbij voor het gebied van de lage toonhoogten het ene detectie-algoritme en voor het gebied van hoge toonhoogten het andere algoritme betrouwbaarder is. Dit doel wordt volgens de uitvinding daardoor bereikt, dat in een eerste elementaire toonhoogtemeter het amplitudespectrum van het spraaksegment wordt bepaald en significante piekposities daarin bepaald worden, dat in een tweede elementaire toonhoogtemeter de autocorrelatiefunctie wordt bepaald en significante piekposities daarin bepaald worden, en dat de significante piekposities van het amplitudespectrum resp. de significante piekposities van de autocorrelatiefunctie de ingangsgegevens vormen van de werkwijze, omvattende de navolgende stappen:

- het kiezen van een waarde voor de toonhoogte resp. periode en het bepalen van een reeks opeenvolgende gehele veelvouden van deze waarde en het bepalen van intervallen rondom deze waarde en de veelvouden daarvan, welke intervallen een masker definiëren met openingen ter plaatse van een interval, aan welke openingen harmonische nummers zijn toegevoegd corresponderende met de vermenigvuldigde factoren in de genoemde veelvouden;
- het berekenen van een kwaliteitsgetal overeenkomstig een criterium dat de mate aangeeft waarin de significante piekposities en de openingen van het masker bij elkaar passen;
- het herhalen van de voorafgaande stappen voor opeenvolgende hogere waarden van de toonhoogte resp. periode tot een bepaalde hoogste waarde, waardoor een reeks bij deze waarden van de toonhoogte resp. periode behorende kwaliteitsgetallen verkregen wordt;
- het selecteren van een drietal waarden van de toonhoogte resp. periode met de hoogste kwaliteitsgetallen;
- het omzetten van de waarden voor de periode in waarden voor de toonhoogte;
- het combineren van de zo gevonden waarden voor de toonhoogte met de bijbehorende kwaliteitsgetallen tot een schatting van de toonhoogte die het meest waarschijnlijk is.

8400552

Bij de combinatie van gegevens kunnen ook nog andere gegevens mee in rekening genomen worden, bijvoorbeeld meetgegevens uit het recente verleden om zodoende ook tijdcontinuïteit van de bepaling van de toonhoogte te garanderen.

5

C. Korte beschrijving van de figuren

Fig. 1: Blokschema van het uitvoeringsvoorbeeld.

- 10 Fig. 2: Blokschema van een procedure die meerdere malen wordt toegepast en die ten doel heeft een harmonische relatie te ontdekken tussen een reeks getallen aan de ingang.

Fig. 3: Blokschema voor het bepalen van significante piekposities in
15 het amplitudespectrum.

Fig. 4: Uitgewerkt stroomdiagram van de procedure voor het bepalen van een drietal f_0 -schattingen met de hoogste kwaliteitsgetallen, gebaseerd op de significante piekposities in het amplitudespectrum.

20

Fig. 5: Blokschema voor het bepalen van significante piekposities in de genormaliseerde autocorrelatiefunctie.

Fig. 6: Uitgewerkt stroomdiagram van de procedure voor het bepalen van
25 een drietal f_0 -schattingen met de hoogste kwaliteitsgetallen, gebaseerd op de significante piekposities in de genormaliseerde autocorrelatiefunctie.

Fig. 7: Stroomdiagram van de combinatieprocedure die de gegevens combi-
30 neert tot een meer betrouwbare schatting van de toonhoogte.

D. Referenties

1. L.R. Rabiner et al., 'A semi-automatic pitch detector (SAPD)', IEEE
35 Transactions on acoustics, speech and signal processing, Vol. ASSP-23, No. 6, December 1975, pp 570-574.
2. L.R. Rabiner, 'On the use of autocorrelation analysis for pitch detection', IEEE Transactions on acoustics, speech and signal pro-

8400552

cessing, Vol. ASSP-25, No. 1, February 1977, pp 24-33.

3. Nederlandse octrooi-aanvraag 78 12 151 (PHN 9313).

E. Uitvoeringsvoorbeeld

5

Het in Fig. 1 aangegeven spraakanalysesysteem dient er toe de toonhoogte van spraaksignalen in een gebied van 50 Hz tot 500 Hz te bepalen. Dit doel wordt gerealiseerd in een spraakanalysesysteem van het onderhavige type door:

- 10 - uit te gaan van een spraaksegment van 40 ms lang, zoals is aangeduid in blok 10;
- het bepalen van het amplitudespectrum van dit segment door het toepassen van een venster in blok 11 en een Fourier-transformatie in blok 12;
- 15 - het bepalen van significante piekposities in dit amplitudespectrum zoals aangeduid in blok 13;
- het nagaan of de gevonden piekposities passen bij een harmonische reeks in blok 14 met het inschrift: HRMSV. De functie van blok 14 wordt aangeduid als harmonische zeef en omvat de volgende stappen:
- 20 * het kiezen van een waarde voor de toonhoogte en het bepalen van een reeks opeenvolgende gehele veelvouden van deze waarde en het bepalen van intervallen rondom deze waarde en de veelvouden daarvan, welke intervallen een masker definiëren met openingen ter plaatse van een interval, aan welke openingen harmonische nummers zijn toe-
- 25 gevoegd corresponderende met de vermenigvuldigde factoren in de genoemde veelvouden;
- * het berekenen van een kwaliteitsgetal overeenkomstig een criterium dat de mate aangeeft waarin de significante piekposities en de openingen van het masker bij elkaar passen;
- 30 * het herhalen van de voorafgaande stappen voor opeenvolgende hogere waarden van de toonhoogte tot een bepaalde hoogste waarde, waardoor een reeks bij deze waarden van de toonhoogte behorende kwaliteitsgetallen verkregen wordt;
- * het selecteren van een drietal waarden van de toonhoogte met de
- 35 hoogste kwaliteitsgetallen.
- het bepalen van significante piekposities in de autocorrelatiefunctie (blok 15) van datzelfde spraaksegment in blok 16;
- het nagaan of de gevonden piekposities passen bij een harmonische

8400552

reeks zoals aangegeven in blok 17, die qua werking gelijk is aan blok 14. Dit geschiedt door:

- 5 * het kiezen van een waarde voor de periode en het bepalen van een reeks opeenvolgende gehele veelvouden van deze waarde en het bepalen van intervallen rondom deze waarde en de veelvouden daarvan, welke intervallen een masker definiëren met openingen ter plaatse van een interval, aan welke openingen harmonische nummers zijn toegevoegd corresponderende met de vermenigvuldigingsfactoren in de genoemde veelvouden;
- 10 * het berekenen van een kwaliteitsgetal overeenkomstig een criterium dat de mate aangeeft waarin de significante piekposities en de openingen van het masker bij elkaar passen;
- * het herhalen van de voorafgaande stappen voor opeenvolgende hogere waarden van de periode tot een bepaalde hoogste waarde, waardoor
15 een reeks bij deze waarden van de toonhoogte behorende kwaliteitsgetallen verkregen wordt;
- * het selecteren van een drietal waarden van de periode met de hoogste kwaliteitsgetallen;
- het omzetten van de waarden voor de periode in waarden voor de toon-
20 hoogte;
- het combineren van de zo gevonden waarden voor de toonhoogte met de bijbehorende kwaliteitsgetallen tot een schatting van de toonhoogte die het meest waarschijnlijk is aangeduid met blok 18.

- 25 In het hier beschreven spraakanalysesysteem vormt de zogenoemde harmonische zeef, aangeduid met blok 14 en 17 in Fig. 1 een belangrijk onderdeel.

De werking van de harmonische zeef is nader aangeduid in Fig. 2 die
30 opereert op significante piekposities $p(i)$ hetzij frequenties (blok 14) hetzij perioden (blok 17). De beschrijving zal plaatsvinden met betrekking tot blok 14 in termen van frequenties (toonhoogtes), wanneer deze veranderd worden in perioden dan heeft de beschrijving betrekking op blok 17. In dit proces wordt eerst een waarde F_s voor de toonhoogte
35 aangenomen, zoals gerepresenteerd in blok 19.

Rondom deze beginwaarde en een aantal opeenvolgende gehele veelvouden daarvan worden intervallen gedefinieerd. Deze intervallen worden beschouwd als openingen in een masker in de zin dat een getalwaarde welke

coïncideert met een opening door het masker zal worden doorgelaten. In deze opvatting functioneert het masker als een soort zeef voor getalwaarden. Deze operaties worden gerepresenteerd door blok 20 met de inscriptie MSK.

- 5 Aan de openingen van een masker zijn nummers toegevoegd, welke worden aangeduid als harmonische nummers en welke corresponderen met de vermenigvuldigfactoren van de betreffende veelvouden van de gekozen waarde van de toonhoogte.

- In een volgende bewerking wordt bepaald in welke mate de significante
10 piekposities $p(i)$ en de openingen van het masker bij elkaar passen. Wanneer weinig significante piekposities door het masker worden doorgelaten dan is er duidelijk een slechte aanpassing. Wanneer anderzijds veel van de piekposities worden doorgelaten maar veel openingen in het masker geen significante piekposities doorlaten omdat deze op die
15 plaats niet aanwezig zijn, dan is er eveneens sprake van een slechte aanpassing.

- Het is mogelijk een geschikt criterium te vinden om de mate van aanpassing tot uitdrukking te brengen in een kwaliteitsgetal, zoals in het navolgende nog zal worden toegelicht. Op dit punt van de beschrijving
20 is het voldoende te melden dat een kwaliteitsgetal wordt berekend voor het masker. Deze operatie wordt gerepresenteerd door blok 21, met de inscriptie QLT.

- In de beslissingsruut 22 wordt nagegaan of de waarde F_s welke voor de toonhoogte gekozen is kleiner is dan een bepaalde maximale waarde:
25 $F_s < M_x$. Wanneer dit zo is, dan wordt de Y-tak van ruit 22 gevolgd, waardoor een lus 23 ontstaat naar blok 24. In deze lus wordt de waarde van F_s op een bepaalde wijze verhoogd; met een bepaald bedrag of een bepaald percentage. Deze functie wordt gerepresenteerd door blok 24 met de inscriptie NCR F_s .

- 30 Het gevolg van de aanwezigheid van beslissingsruut 22 is dat de procedures welke worden gerepresenteerd door de blokken 20 en 21 voortdurend worden herhaald voor steeds nieuwe waarden van F_s totdat F_s de maximale waarde M_x bereikt. Wanneer dit het geval is dan wordt de N-tak gevolgd en wordt lus 23 verlaten.

- 35 De volgende procedure in het huidige spraakanalysesysteem bestaat dan in het bepalen van een drietal waarden van F_s waarvan de kwaliteitsgetallen de hoogste waarden hebben. Dit geschiedt in blok 25 met het inschrift SLCT F_s .

8400552

In het onderhavige spraakanalysesysteem wordt vervolgens een nauwkeurige schatting gemaakt van de mogelijke toonhoogtes uitgaande van de drie geselecteerde waarden F_s . Deze laatste stap in de procedure voor het bepalen van de toonhoogte wordt gerepresenteerd door blok 26 met het
 5 inschrift STM EP(1,2,3) waarvan de uitgangstak de drie geschatte waarden EP(1,2,3) van de toonhoogte levert. In dit blok 26 worden de harmonische nummers van de openingen van het referentiemasker toegevoegd aan de met deze openingen coïnciderende significante piekposities $p(i)$ en ieder van deze piekposities $p(i)$ zal dan een harmonisch nummer n_i krij-
 10 gen welke de plaats van de piekposities bepaalt in een reeks harmonischen van dezelfde grondtoon. Een goede schatting van F_0 : $\hat{F}_0$ kan gedefinieerd worden als de waarde waarvoor de afwijkingen tussen de laatstgenoemde significante piekposities $p(i)$ en de overeenkomstige veelvouden $n_i \cdot \hat{F}_0$ van de waarschijnlijke waarde zo klein mogelijk zijn. Wanneer
 15 voor het bepalen van de afwijkingen een m.s.e.-criterium (mean-square-error) wordt toegepast dan laat $\hat{F}_0$ zich berekenen door de uitdrukking:

$$\hat{F}_0 = \sum_{i=1}^K p(i) \cdot n_i / \sum_{i=1}^K n_i^2 \quad (1)$$

20 De sommatie in deze uitdrukking strekt zich uit over alle significante piekposities welke coïncideren met een opening van het referentiemasker waarvan het aantal door K wordt gerepresenteerd. Overigens vormt de waarde van de toonhoogte welke behoort bij het referentiemasker reeds een eerste schatting van de gezochte toonhoogte.

25 In Fig. 3 is in meer detail de procedure geïllustreerd voor het verkrijgen van de waarden van de significante piekposities in het frequentiedomein.

Van het bemonsterde spraaksignaal worden tijdsegmenten genomen met een
 30 duur van 40 ms. Deze functie wordt gerepresenteerd door blok 27, met de inscriptie 40 ms. De volgende bewerking is het vermenigvuldigen van het spraaksignaalsegment met een zogenaamd 'Hamming window', welke functie wordt gerepresenteerd door blok 28 met de inscriptie WNDW.

De monsters van het spraaksignaalsegment worden vervolgens onderworpen
 35 aan een discrete Fourier-transformatie met 256 punten zoals gerepresenteerd door blok 29 met de inscriptie DFT.

In de volgende operatie worden de amplituden van 128 spectrumcomponenten bepaald uit de 256 reële en imaginaire waarden welke door de DFT

geleverd worden. Uit deze spectrumcomponenten worden de significante piekposities $PF(i)$ afgeleid welke de plaatsen van de pieken in het spectrum representeren.

- Sommige operaties van het onderhavige spraakanalysesysteem kunnen worden geïmplementeerd in de software van een general-purpose computer. Andere operaties kunnen versneld worden door toepassing van externe hardware.

Vanaf blok 30 wordt de procedure geïmplementeerd door de software van een general-purpose computer.

- 10 De computer ontvangt als ingangsgegevens de componenten $AF(r)$, $r=1, \dots, 128$ van het amplitudespectrum zoals gerepresenteerd door blok 31. Als beginwaarden voor de routine worden $r=2$ en $NTOP=0$ gezet. Deze functie wordt gerepresenteerd door blok 32. $NTOP$ is een variabele, die het aantal gevonden locale maxima telt.

- 15 Te beginnen met spectrumcomponent $AF(2)$ wordt in beslissingsruit 33 onderzocht of de spectrumcomponent $AF(2)$ groter is dan een drempelwaarde THF . De N-tak van ruit 33 voert naar blok 39 welke aangeeft dat r met één verhoogd wordt. Daarna wordt in beslissingsruit 40 onderzocht of r groter of gelijk is geworden aan 127. Zolang dit niet het geval is
20 wordt een lus 41 gevormd naar blok 33. De functie van blok 33 wordt dan herhaald met een nieuwe waarde van r .

- De Y-tak van beslissingsruit 33 voert naar beslissingsruit 34 waarin wordt onderzocht of de spectrumcomponent $AF(2)$ groter of gelijk is aan de voorafgaande spectrumcomponent $AF(1)$ en of spectrumcomponent $AF(2)$
25 groter is dan de volgende spectrumcomponent $AF(3)$. Deze functie wordt gerepresenteerd door beslissingsruit 34. Wanneer de spectrumcomponent een lokaal maximum vormt, dan wordt de Y-tak van ruit 36 gevolgd.

De N-tak van ruit 34 voert naar blok 39 welke aangeeft dat r met één verhoogd wordt zolang de nieuwe waarde van r kleiner is dan 127.

- 30 De drempelwaarde THF wordt in de eerste plaats gevormd door een absolute waarde welke wordt bepaald door het niveau van de ruis welke een gevolg is van de kwantisering en de 'Hamming window'.

- In de tweede plaats kan een deel van de drempelwaarde THF variabel zijn om rekening te houden met het maskeren van een spectrumcomponent door
35 de naburige spectrumcomponenten wanneer deze een veel grotere amplitude hebben. Dit effect treedt op bij het menselijke gehoor en is daar een belangrijke factor bij de waarneming van de toonhoogte.

Wanneer de Y-tak van beslissingsruit 34 wordt gevolgd, dan wordt een

operatie uitgevoerd om de amplitude en de frequentie van het locale maximum van het amplitudespectrum te bepalen. Hiervoor wordt gebruik gemaakt van interpolatie tussen de waarden $AF(r-1)$, $AF(r)$ en $AF(r+1)$ met een tweedegraads polynoom (parabolische interpolatie). Deze functie
5 wordt gerepresenteerd door blok 36 met het inschrift INTRP. In blok 37 wordt nu het aantal locale maxima NTOP met één verhoogd.

Het zoeken naar locale maxima van het amplitudespectrum wordt voortgezet tot maximaal zes significante piekposities $PF(i)$ bepaald zijn. Wanneer dit het geval is, dan wordt de Y-tak van beslissingsruit 38 actief
10 en worden de significante piekposities $PF(i)$ uitgevoerd (blok 42).

De significante piekposities $PF(i)$ welke worden geleverd door de routine volgens Fig. 3 vormen de ingangsgegevens voor de routine volgens Fig. 4A en 4B. Deze figuren dienen op de aangegeven wijze onder elkaar te worden geplaatst.

15

Fig. 4A en 4B tonen het stroomdiagram van een programma voor het bepalen van een drietal waarschijnlijke waarden van de toonhoogte gebruikmakende van het concept van de maskers.

Het programma ontvangt als ingangsgegevens de significante piekposities
20 $PF(i)$, $i=1, \dots, N$, zoals geïllustreerd in blok 43. Deze worden alternatief als componenten aangeduid.

In het begin worden een drietal f_0 -schattingen $f_0(j)$, $j=1, 2, 3$ met bijbehorende kwaliteitsgetallen $q(j)$ op nul gezet (blok 44).

Wanneer het aantal aangeboden componenten kleiner is dan één (ruit 45),
25 dan wordt de routine verlaten en de waarden $f_0(i)=0$ uitgevoerd (blok 46).

Wordt één of meer componenten ingevoerd dan wordt de routine verder gevolgd via de N-tak van de beslissingsruit 45.

Als voorbereiding wordt de variabele l welke het nummer van het masker
30 aangeeft op één gezet en wordt de bij dit masker behorende toonhoogte f_{01} op 50 Hz ingesteld (blok 47). Daarna worden enkele variabelen op een beginwaarde gezet (blok 48).

In de volgende procedure (blok 49) wordt te beginnen bij de eerste component $PF(1)$ een schatting gemaakt van het bij de component $PF(1)$ behorende harmonische nummer m_{1k} en wordt deze waarde afgerond naar het
35 meest nabij gelegen gehele getal m_{1k} .

Wanneer m_{1k} groter is dan 11 (beslissingsruit 50), dan wordt een groot deel van het programma overgeslagen, omdat in het onderhavige spraak-

analysesysteem harmonischen met een hoger nummer dan 11 niet bij de toonhoogtebepaling worden betrokken.

Vervolgens wordt nagegaan of m_{1k} de waarde nul heeft (beslissingsruit 52). Is dit niet het geval, dan wordt nagegaan of de component $PF(n)$ in
5 een opening van het masker met toonhoogte f_{01} valt. Wanneer de relatieve afwijking van $PF(n)$ ten opzichte van de meest nabij gelegen harmonische van de grondtoon f_{01} kleiner is dan een bepaald percentage, in het huidige systeem 5%, dan wordt $PF(n)$ geacht in de opening te liggen (beslissingsruit 54).

10 Wanneer de component $PF(n)$ in een opening van het masker ligt, dan wordt de N-tak van beslissingsruit 54 actief.

De volgende operatie betreft nu het geval dat voor m_{1k} dezelfde waarde wordt gevonden als de waarde m_{1K} ($K+1=k$) welke de vorige keer is bepaald. In dit geval liggen er twee componenten in dezelfde opening van
15 het masker. Het onderhavige spraakanalysesysteem accepteert alleen de component welke het dichtst bij het midden van de opening ligt en telt de andere component niet mee.

De variabele K telt het aantal van de componenten welke in een opening liggen. Wanneer m_{1k} groter is dan m_{1K} (beslissingsruit 55) dan wordt
20 daarna K met één verhoogd (blok 58).

Wanneer echter m_{1k} niet groter is dan m_{1K} dan wordt bepaald voor welke van de waarden m_{1k} en m_{1K} de kleinste relatieve afwijking optreedt ten opzichte van het midden van de opening (beslissingsruit 56). Wanneer dit het geval is voor m_{1k} dan wordt $\hat{m}_{1K}$ gelijk gesteld aan $\hat{m}_{1k}$
25 (blok 57). In het andere geval wordt $\hat{m}_{1K}$ niet veranderd. In beide gevallen wordt K niet verhoogd.

Wanneer het programma de Y-tak van beslissingsruit 52, de Y-tak van beslissingsruit 54 of de N-tak van beslissingsruit 56 volgt of na de operaties van de blokken 57 of 58, wordt de waarde van n met één verhoogd
30 (blok 59). De variabele n telt de aangeboden componenten $PF(i)$ en wanneer n niet groter is dan het totaal aantal aangeboden componenten (beslissingsruit 60), dan wordt de lus 61 binnengetreden.

De beschreven routine begint dan opnieuw bij blok 49 voor een nieuwe waarde van n. Op deze wijze wordt de routine herhaald voor alle N componenten $PF(i)$.
35

Wanneer n groter wordt dan N, dan wordt de Y-tak van beslissingsruit 60 gevolgd. Hierna wordt geregistreerd dat voor het masker met index 1 het aantal in aanmerking genomen componenten N_1 gelijk is aan N (blok 62).

8400552

Wanneer het programma de Y-tak van beslissingsruut 50 volgt dan wordt N_1 gelijk gesteld aan n (blok 63). Componenten $PF(i)$ met een hogere indexwaarde hebben een geschat harmonisch nummer dat groter is dan 11 en worden niet in aanmerking genomen bij de toonhoogte-bepaling. Een masker heeft in het huidige spraakanalysesysteem 11 openingen en componenten $PF(i)$ welke buiten het masker liggen, doen niet mee bij de bepaling van de toonhoogte.

De volgende procedure heeft betrekking op de berekening van een kwaliteitsgetal Q waarmee wordt aangegeven de mate waarin de componenten $PF(i)$ en de openingen van het masker op elkaar passen.

Een kwaliteitsgetal kan worden afgeleid door de reeks aangeboden componenten $PF(i)$ en de reeks openingen van een masker te beschouwen als vectoren in een meerdimensionale ruimte. De afstand tussen de vectoren geeft dan aan hoe goed de componenten $PF(i)$ en het masker op elkaar passen. Het kwaliteitsgetal kan dan worden berekend als één gedeeld door de afstand. In de plaats van de afstand kan ook iedere andere uitdrukking genomen worden welke minimaal is als de afstand minimaal is en omgekeerd.

Het kan op elementaire wijze worden aangetoond dat de afstand D kan worden uitgedrukt door:

$$D = \sqrt{N + M - 2K} \quad (2)$$

waarin N het aantal componenten $PF(i)$, M het aantal openingen van het masker en K het aantal van de componenten $PF(i)$ voorstelt welke liggen in de openingen van het masker.

Het kwaliteitsgetal Q kan worden uitgedrukt als:

$$Q = \frac{1}{D^2} = \frac{1}{N + M - 2K} \quad (3)$$

De afstand D kan worden genormaliseerd door deze te delen door de lengte van de eenheidsvector:

$$E = \sqrt{N + M - K} \quad (4)$$

Dit zou resulteren in het kwaliteitsgetal:

8400552

$$Q = \frac{E^2}{D^2} = \frac{N + M - K}{N + M - 2K} \quad (5)$$

Na elementaire bewerkingen kan worden aangetoond dat Q volgens uitdruk-
 5 king (5) maximaal is wanneer Q' volgens de uitdrukking:

$$Q' = \frac{K}{N + M} \quad (6)$$

10 maximaal is.

In het kwaliteitsgetal wordt bij voorkeur tot uitdrukking gebracht dat
 de berekening des te betrouwbaarder is naarmate het aantal componenten
 dat binnen het masker valt groter is. Om dit te bereiken wordt gebruik
 gemaakt van kwaliteitsmaat Q'' waarvoor dan geldt:

$$15 \quad Q'' = \frac{K^2}{N + M} \quad (7)$$

In het gebruikte systeem voor het vinden van de significante piekposi-
 20 ties $PF(i)$ wordt het zoeken gestopt als 6 piekposities zijn gevonden
 (beslissingsruit 38 in Fig. 2). De meest ideale meting is die waarbij
 deze 6 piekposities samenvallen met de eerste zes maskergaten zodat men
 voor het kwaliteitsgetal Q'' de waarde 3 vindt.

Het heeft voordelen het kwaliteitsgetal Q'' te normaliseren met deze
 25 hoogst bereikbare waarde zodat het nieuwe kwaliteitsgetal Q_n wordt.

$$Q_n = \frac{Q''}{3} = \frac{K^2}{3(N + M)} \quad (8)$$

30 Dit kwaliteitsgetal bereikt in het ideale geval de waarde 1 en bereikt
 een lagere waarde in alle andere niet ideale situaties.

Componenten $PF(i)$ welke buiten het masker vallen dragen niet bij tot de
 waarde van K hoewel ze wel een harmonisch verband kunnen hebben met de
 grondtoon van het masker. Een beter bruikbaar kwaliteitsgetal zal wor-
 35 den verkregen wanneer in de uitdrukkingen voor Q de grootheid N wordt
 vervangen door N_1 , welke het aantal componenten aangeeft, dat binnen
 het bereik van het masker ligt.

Het kan voorkomen dat openingen van het masker buiten het bereik van de

8400552

aangeboden componenten vallen en daarom geen component doorlaten. Het kwaliteitsgetal kan hiervoor worden gecorrigeerd door in de uitdrukking voor Q de grootheid M te vervangen door m_{1K} , welke het hoogste nummer is van de openingen welke een component doorlaten.

5

In de procedure volgens Fig. 4A en 4B wordt in blok 63 de kwaliteitsmaat Q_n volgens betrekking (8) uitgerekend, en in blok 64 wordt de nauwkeurige schatting van de eventuele toonhoogte uitgerekend volgens de uitdrukking (1).

10 In blok 65 wordt de waarde van 1 met één opgehoogd en er wordt een nieuwe waarde van f_{01} bepaald, die 3% hoger is dan de vorige waarde.

In beslissingsruit 66 wordt nagegaan of 1 groter is dan een grenswaarde L . Deze grenswaarde in het huidige spraakanalysesysteem op 80 gesteld. Is dit niet het geval dan wordt de ruit 66 aan de N-tak verlaten en

15 wordt lus 67 binnengegaan waarna het hele proces van het zoeken opnieuw begint. Is echter de grenswaarde L overschreden dan wordt de ruit 66 langs de Y-tak verlaten en worden in blok 68 de drie grootste kwaliteitsgetallen met de bijbehorende schattingen van de toonhoogte gezocht die dan aan de uitgang van de operatie in blok 69 beschikbaar zijn.

20

In Fig. 5 is in meer detail aangegeven de procedure voor het verkrijgen van de waarden van de significante piekposities in het tijddomein. Aan deze procedure ligt hetzelfde spraaksegment van 40 ms (blok 70) ten grondslag als in Fig. 3 (blok 27). Van dit signaal wordt nu in blok 71
25 aangeduid met NRG de energie berekend. Deze energie E is gedefinieerd door:

$$E = \frac{1}{N} \sum_{K=1}^N S_K^2 \quad (N = 100) \quad (9)$$

30

Van het spraaksegment wordt nu de genormaliseerde autocorrelatiefunctie berekend in blok 72 volgens de uitdrukking:

$$AT(j) = \frac{\frac{1}{N-j} \sum_{K=1}^{N-j} S_K \cdot S_{K+j}}{E} \quad (10)$$

voor $j=1, \dots, 80$.

8400552

Deze functie wordt voorgesteld in blok 73, waarin de variabele j is vervangen door r . Als beginwaarden voor de nu volgende routine worden in blok 74 gezet: $r=2$ en $NTOP=0$.

Te beginnen met de autocorrelatiecoëfficiënt $AT(2)$ wordt in beslissingsruit 75 onderzocht of de autocorrelatiecoëfficiënt $AT(2)$ groter is dan een drempelwaarde THA . De N-tak van ruit 75 voert naar blok 81 welke aangeeft dat r met één wordt verhoogd. Daarna wordt in beslissingsruit 83 onderzocht of r groter of gelijk is geworden aan 79. Zolang dit niet het geval is wordt de lus 82 naar de beslissingsruit 75 gevolgd.

10 De functie van beslissingsruit 75 wordt dan herhaald met een nieuwe waarde van r .

De Y-tak van beslissingsruit 75 voert naar beslissingsruit 76 waarin wordt onderzocht of de autocorrelatiecoëfficiënt groter of gelijk is aan de voorafgaande autocorrelatiecoëfficiënt $AT(1)$ en of autocorrelatiecoëfficiënt $AT(2)$ groter is dan de volgende autocorrelatiecoëfficiënt $AT(3)$. Wanneer de autocorrelatiecoëfficiënt een lokaal maximum vormt, dan wordt de Y-tak van ruit 76 gevolgd. De N-tak van ruit 76 voert naar blok 81, welke aangeeft dat r met één wordt verhoogd. Wanneer de Y-tak van beslissingsruit 76 wordt gevolgd, dan wordt een operatie uitgevoerd om de positie op de tijdsas van het locale maximum van de autocorrelatiefunctie te bepalen. Hiervoor wordt gebruik gemaakt van interpolatie tussen de waarden $AT(r-1)$, $AT(r)$ en $AT(r+1)$ met een tweedegraads polynoom (parabolische interpolatie). Deze functie wordt gerepresenteerd door blok 77 met het inschrift INTRP. In blok 78 wordt nu

15

20

25

het aantal locale maxima $NTOP$ met één verhoogd. Het zoeken naar locale maxima in de autocorrelatiefunctie wordt voortgezet tot maximaal zes significante piekposities $PP(i)$ bepaald zijn.

Wanneer zes significante piekposities zijn gevonden, dan wordt de Y-tak van de beslissingsruit 80 actief en dan worden de significante piekposities uitgevoerd (blok 84).

30

De significante piekposities $PP(i)$ welke worden geleverd door de routine volgens Fig. 5 vormen de ingangsgegevens voor de routine volgens Fig. 6a en 6b. Deze figuren dienen op de aangegeven wijze onder elkaar te worden geplaatst.

35

Fig. 6A en 6B tonen het stroomdiagram van een procedure voor het bepalen van een drietal waarschijnlijke waarden van de toonhoogte gebruikmakende van het concept van de maskers. Het concept van de maskers

8400552

wordt nu toegepast op de significante piekposities $PP(i)$ die liggen in het tijddomein en stellen derhalve periodeduren voor.

Het programma ontvangt als ingangsgegevens de significante piekposities $PP(i)$, $i=1\dots N$, zoals geïllustreerd in blok 90. Deze worden alternatief
5 als componenten aangeduid. In het begin worden een drietal t_o -schattingen $t_o(i)$, $i=1,2,3$ met bijbehorende kwaliteitsgetallen $s(i)$ op nul gezet (blok 91). Wanneer het aantal aangeboden componenten kleiner is dan één (ruit 92) dan wordt de routine via de Y-tak van ruit 92 verlaten en worden de waarden $t_o(i)=0$ uitgevoerd (blok 93). Worden één of meer com-
10 ponenten ingevoerd dan wordt de routine verder gevolgd via de N-tak van ruit 92.

Als voorbereiding wordt de variabele l welke het nummer van het masker aangeeft op één gezet en wordt de bij dit masker behorende periodeduur t_{o1} op 2 ms ingesteld (blok 94). In de volgende operatie (blok 95)
15 worden enkele variabelen op hun beginwaarde gezet. In blok 96 wordt te beginnen bij de eerste component $PP(1)$ een schatting gemaakt van het bij de component $PP(1)$ behorende harmonische nummer m_{1k} en wordt deze waarde afgerond naar het meest nabij gelegen gehele getal m_{1k} . Wanneer m_{1k} groter is dan 11 (beslissingsruit 97) dan wordt een groot gedeelte
20 van de procedure via de lus 98 overgeslagen, omdat in het onderhavige spraakanalysesysteem een harmonische relatie met een hoger nummer dan 11 niet bij de toonhoogtebepaling wordt betrokken.

Vervolgens wordt nagegaan of m_{1k} de waarde nul heeft (in beslissingsruit 99). Is dit niet het geval dan wordt ruit 99 via de N-tak verlaten
25 en wordt nagegaan of de component $PP(n)$ in een opening van het masker met periode t_{o1} valt. Wanneer de relatieve afwijking van $PP(n)$ ten opzichte van het meest nabij gelegen veelvoud van de grondperiode t_{o1} kleiner is dan een bepaald percentage, in het huidige systeem 5%, dan wordt $PP(n)$ geacht in de opening te liggen (beslissingsruit 101). Wan-
30 neer de component $PP(n)$ in een opening van het masker ligt, dan wordt de N-tak van beslissingsruit 101 actief.

De volgende operatie betreft het geval dat voor m_{1k} dezelfde waarde wordt gevonden als de waarde m_{1K} ($K+1=k$) welke de vorige keer is bepaald. In dit geval liggen er twee componenten in dezelfde opening van
35 het masker.

Het onderhavige spraakanalysesysteem accepteert alleen de component welke het dichtst bij het midden van de opening ligt en telt de andere component niet mee. De variabele K telt het aantal van de componenten

welke in een opening liggen. Wanneer m_{1k} groter is dan m_{1K} (beslissingsruit 102) dan wordt daarna K met één verhoogd (blok 105). Wanneer echter m_{1k} niet groter is dan m_{1K} dan wordt ruit 102 via de N-tak verlaten en wordt bepaald voor welke van de waarden m_{1k} en m_{1K} de kleinste afwijking optreedt ten opzichte van het midden van de opening (beslissingsruit 103). Wanneer dit het geval is voor m_{1k} , dan wordt $\hat{m}_{1K}$ gelijk gesteld aan $\hat{m}_{1k}$ (blok 104). In het andere geval wordt $\hat{m}_{1K}$ niet veranderd. In beide gevallen wordt K niet verhoogd.

Wanneer het programma de Y-tak van beslissingsruit 99, de Y-tak van beslissingsruit 101 of de N-tak van beslissingsruit 103 volgt of na de operaties van de blokken 105 of 104, wordt de waarde van n met één verhoogd (blok 106).

De variabele n telt de aangeboden componenten PP(n) en wanneer n niet groter is dan het totaal aantal aangeboden componenten (beslissingsruit 107), dan wordt de lus 108 gevolgd. De beschreven routine begint dan opnieuw bij blok 96 voor een nieuwe waarde van n. Op deze wijze wordt de routine herhaald voor alle N componenten PP(i).

Wanneer n groter wordt dan N, dan wordt de Y-tak van beslissingsruit 107 gevolgd. Hierna wordt geregistreerd dat voor het masker met index 1 het aantal in aanmerking genomen componenten N_1 gelijk is aan N (blok 109). Wanneer het programma de Y-tak van beslissingsruit 97 volgt dan wordt N_1 gelijkgesteld aan n (blok 110). Componenten PP(i) met een hogere indexwaarde hebben een geschat harmonisch nummer, dat groter is dan 11 en worden niet in aanmerking genomen bij de toonhoogtebepaling. Een masker heeft in het huidige spraakanalysesysteem 11 openingen en componenten PP(i) welke buiten het masker liggen, doen niet mee met de bepaling van de toonhoogte.

In het blok 111 wordt nu de kwaliteitsmaat uitgerekend volgens betrekking (8) en in blok 112 wordt de nauwkeurige schatting van de eventuele periode uitgerekend volgens de uitdrukking (1).

In blok 113 wordt l met één verhoogd en er wordt een nieuwe waarde van t_{01} berekend, die 3% hoger is dan de vorige. In beslissingsruit 115 wordt nagegaan of l groter is dan een grenswaarde L. Deze grenswaarde is in het onderhavige spraakanalysesysteem op 80 gesteld. Is l niet groter dan L dan wordt ruit 115 via de N-tak verlaten, waarna lus 114 wordt binnengegaan, en het hele proces van het zoeken opnieuw begint. Is echter de grenswaarde L overschreden dan wordt beslissingsruit via de Y-tak verlaten, waarna in blok 116 de drie grootste kwaliteitsgetal-

len $S(K)$ met de bijbehorende periodeschattingen $t_o(k)$ worden gezocht. Deze drie best passende periodeschattingen $t_o(i)$ met bijbehorende kwaliteitsgetallen $s(j)$ zijn nu beschikbaar in blok 117 en worden vervolgens in blok 118 omgezet naar een schatting van de toonhoogte door de
5 inverse van $t_o(j)$ te berekenen.

Er zijn nu een drietal schattingen voor de toonhoogte met bijbehorende kwaliteitsgetallen beschikbaar verkregen uit de toonhoogtemeter werkende in het frequentiedomein, genoemd $f_o(j)$, $j=1,2,3$ zoals aangegeven in
10 blok 69 en een drietal schattingen voor f_o met bijbehorende kwaliteitsgetallen verkregen uit de autocorrelatie-toonhoogtemeter werkende in het tijddomein, genoemd $f_o(i)$, $i=4,5,6$, zoals is aangegeven in blok 119. In de nu volgende combinatieschakeling CMB (blok 18, Fig. 1) worden deze resultaten gecombineerd tot een meer betrouwbare meting van de toon-
15 hoogte.

Bovendien is hier de principiële mogelijkheid aanwezig om meer gegevens dan de hier genoemde te laten meebeslissen over de uiteindelijk aan te wijzen toonhoogte.

Hierbij valt te denken aan een andere nog nader te specificeren toon-
20 hoogtemeter, of aan de toonhoogteschattingen van het vorige meetinterval met gereduceerde kwaliteitsgetallen (gereduceerd om gegevens uit het verleden wat minder te laten wegen bij de bepaling van de huidige toonhoogte) of aan de meetresultaten afgeleid uit het recente verleden (tracking).

25 De combinatieschakeling is weergegeven in Fig. 7 en gaat uit van de gegevens in blok 120, zijnde de zes mogelijke schattingen van de toonhoogte met bijbehorende kwaliteitsgetallen.

In blok 121 wordt de telvariabele m op één gezet en in blok 122 wordt de grootheid $SCR(m)$ op nul gezet. In blok 123 wordt de telvariabele k
30 die in lus 128 actief is, op één gezet. Als de relatieve afwijking tussen de m -de toonhoogteschatting en de k -de toonhoogteschatting kleiner is dan 12,5%, dan wordt de beslissingsruit 124 via de Y-tak verlaten. In blok 125 wordt in dit geval het produkt van de kwaliteitsgetallen van de m -de en de k -de toonhoogteschatting bij $SCR(m)$ opgeteld. Wordt
35 ruit 124 via de N-tak verlaten dan wordt geen bijdrage bij $SCR(m)$ opgeteld en wordt blok 126 binnengegaan waar de variabele k met één wordt verhoogd. In beslissingsruit 127 wordt nagegaan of de variabele k groter is dan 6. Is dit niet het geval dan wordt via de N-tak van ruit 127

8400552

de lus 128 binnengegaan. Is de variabele k groter dan 6 geworden dan wordt beslissingsruit 127 door de Y-tak verlaten, waarna in blok 129 de variabele m wordt verhoogd met één. In beslissingsruit 130 wordt nagegaan of de variabele m groter is dan 6. Is dit niet het geval dan wordt
5 de ruit 130 via de N-tak verlaten en wordt de lus 131 binnengegaan. Is de variabele m groter dan 6 dan wordt de ruit 130 via de Y-tak verlaten. Dan is in $SCR(m)$ voor alle 6 toonhoogteschattingen berekend hoe goed de 6 toonhoogteschattingen met elkaar overeenkomen. In blok 132 wordt nu nog de index j bepaald waarvoor de bijbehorende $SCR(j)$ de
10 hoogste waarde aanneemt. Tenslotte komt dan de toonhoogteschatting $f_0(j)$ als de meest waarschijnlijke beschikbaar in blok 133.

15

20

25

30

35

8400552

Conclusie

1. Systeem voor het analyseren van menselijke spraak voor het bepalen van de toonhoogte van spraaksegmenten onder toepassing van meer dan één toonhoogtedetectie-algoritme met het kenmerk dat in een eerste elementaire toonhoogtemeter het amplitudespectrum van het spraaksegment wordt bepaald en significante piekposities daarin bepaald worden, dat in een tweede elementaire toonhoogtemeter de autocorrelatiefunctie wordt bepaald en significante piekposities daarin bepaald worden, en dat de significante piekposities van het amplitudespectrum resp. de significante piekposities van de autocorrelatiefunctie de ingangsgegevens vormen van de werkwijze omvattende de navolgende stappen:
- het kiezen van een waarde voor de toonhoogte resp. periode en het bepalen van een reeks opeenvolgende gehele veelvouden van deze waarde en het bepalen van intervallen rondom deze waarde en de veelvouden daarvan, welke intervallen een masker definiëren met openingen ter plaatse van een interval, aan welke openingen harmonische nummers zijn toegevoegd corresponderende met de vermenigvuldigde factoren in de genoemde veelvouden.
 - het berekenen van een kwaliteitsgetal overeenkomstig een criterium dat de mate aangeeft waarin de significante piekposities en de openingen van het masker bij elkaar passen.
 - het herhalen van de voorafgaande stappen voor opeenvolgende hogere waarden van de toonhoogte resp. periode tot een bepaalde hoogste waarde, waardoor een reeks bij deze waarden van de toonhoogte resp. periode behorende kwaliteitsgetallen verkregen wordt.
 - het selecteren van een drietal waarden van de toonhoogte resp. periode met de hoogste kwaliteitsgetallen.
 - het omzetten van de waarden voor de periode in waarden voor de toonhoogte.
 - het combineren van de zo gevonden waarden voor de toonhoogte met de bijbehorende kwaliteitsgetallen tot een schatting van de toonhoogte die het meest waarschijnlijk is.

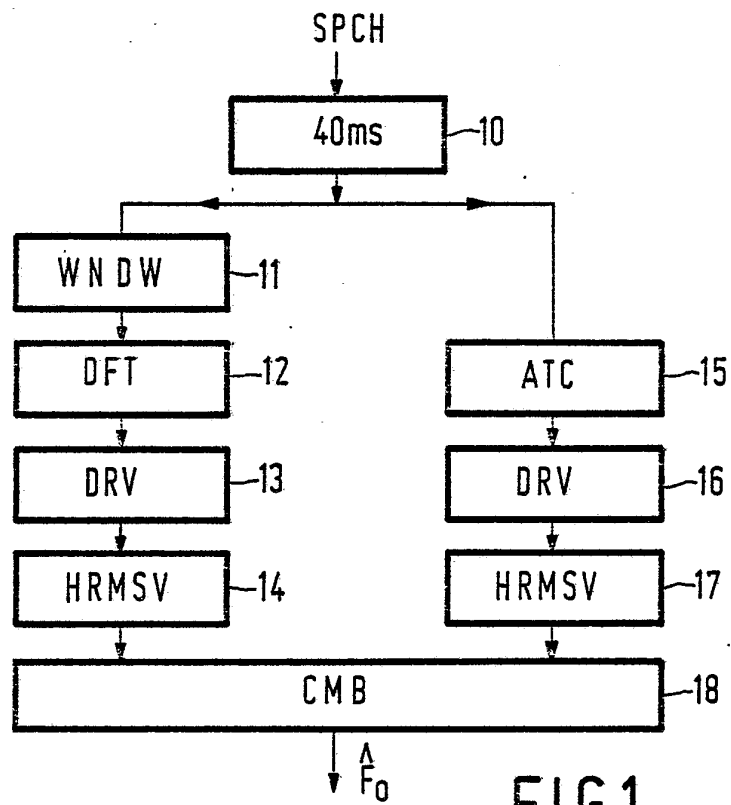


FIG. 1

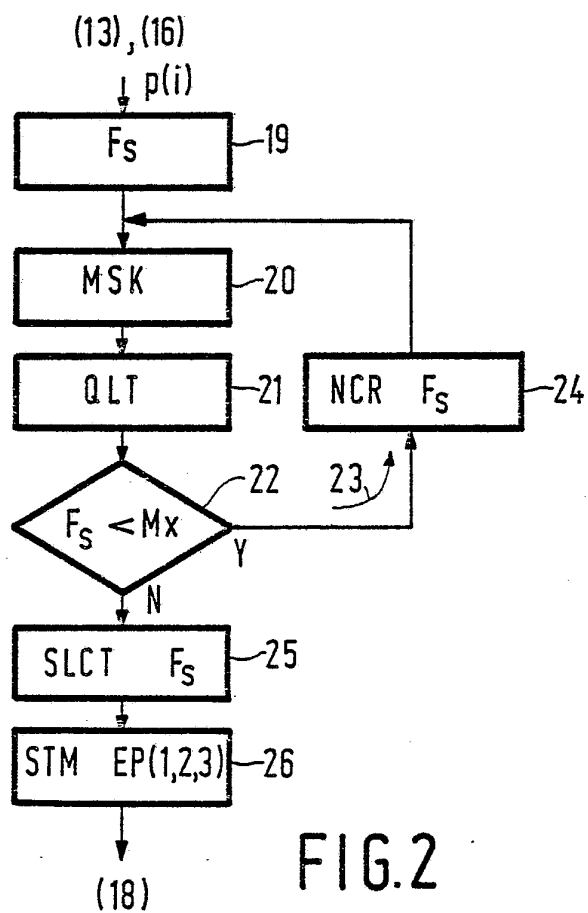


FIG. 2

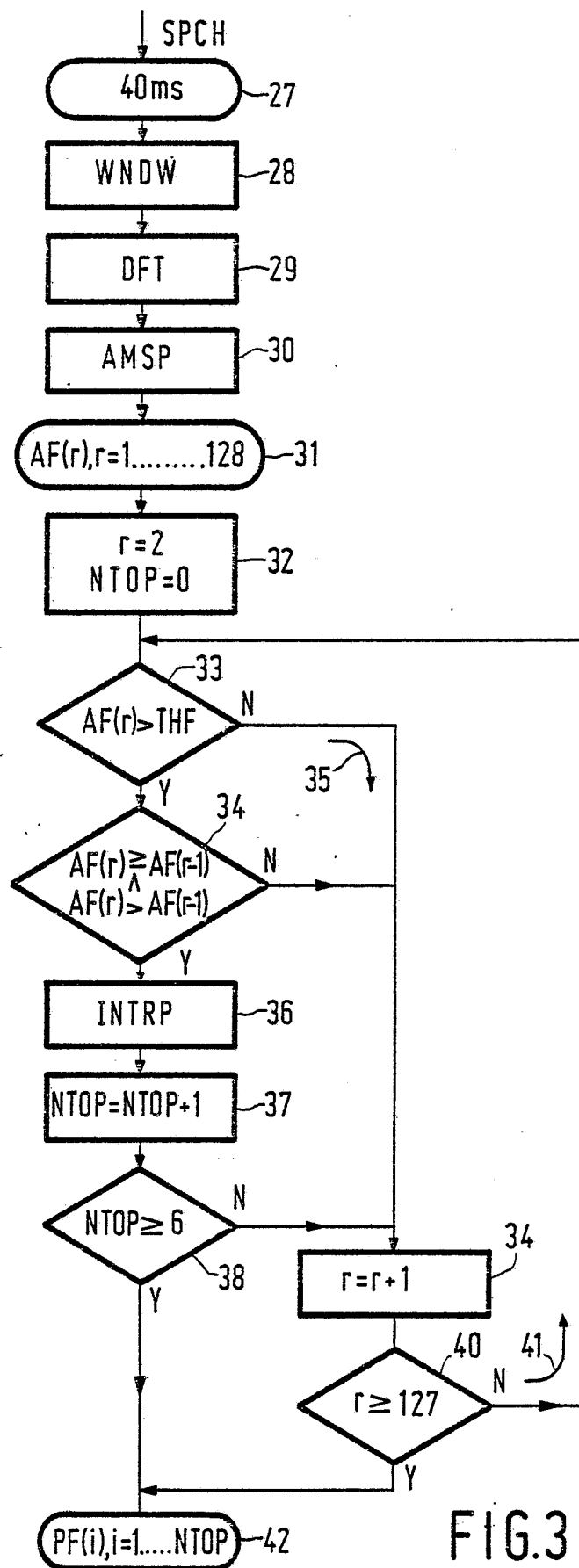
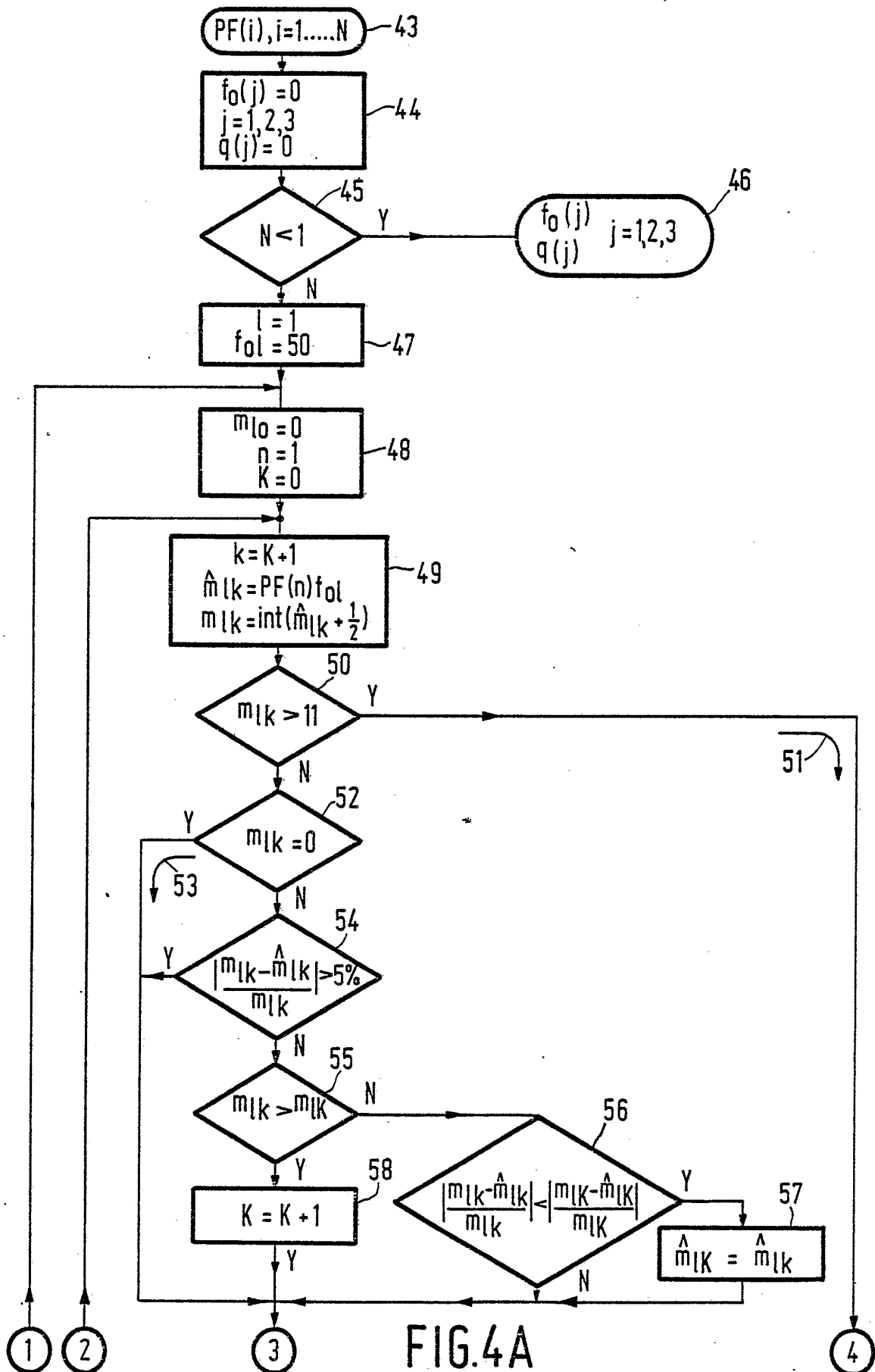


FIG. 3



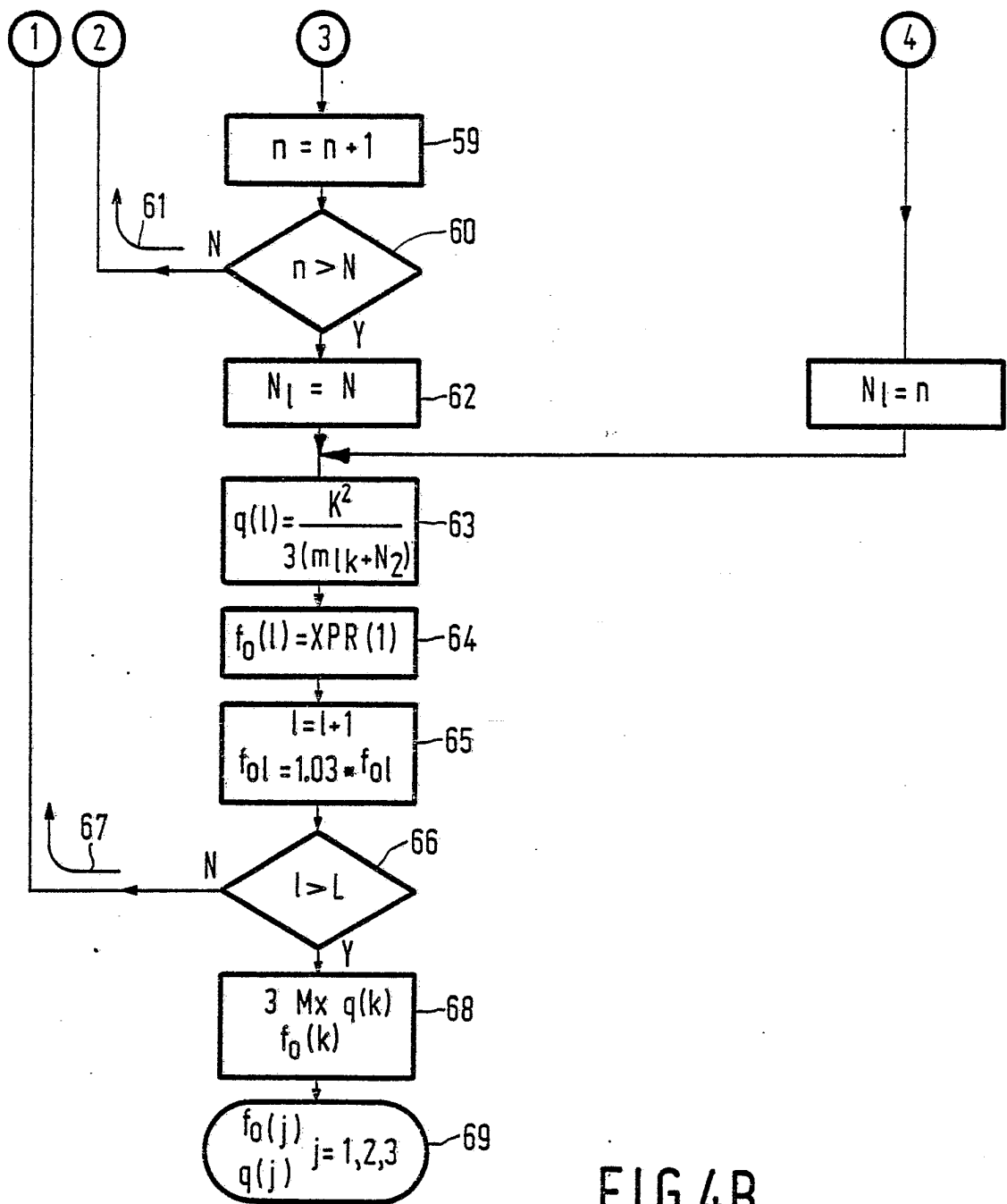
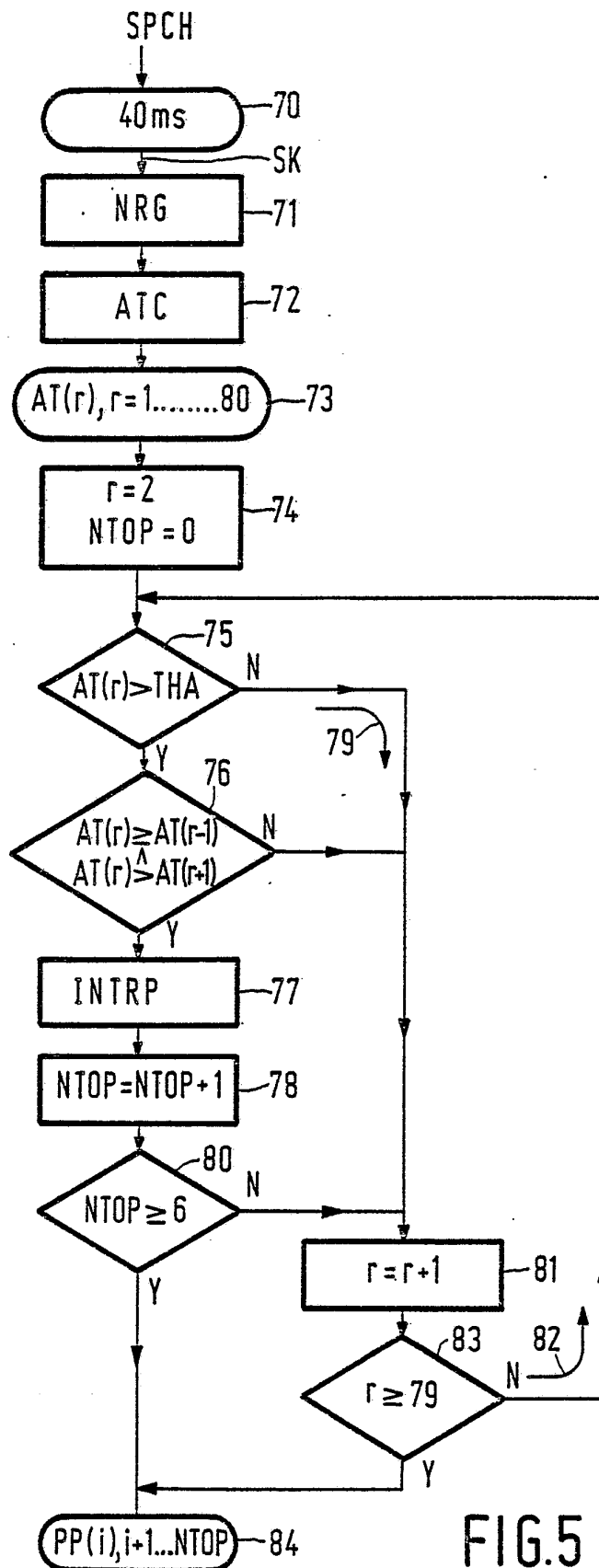
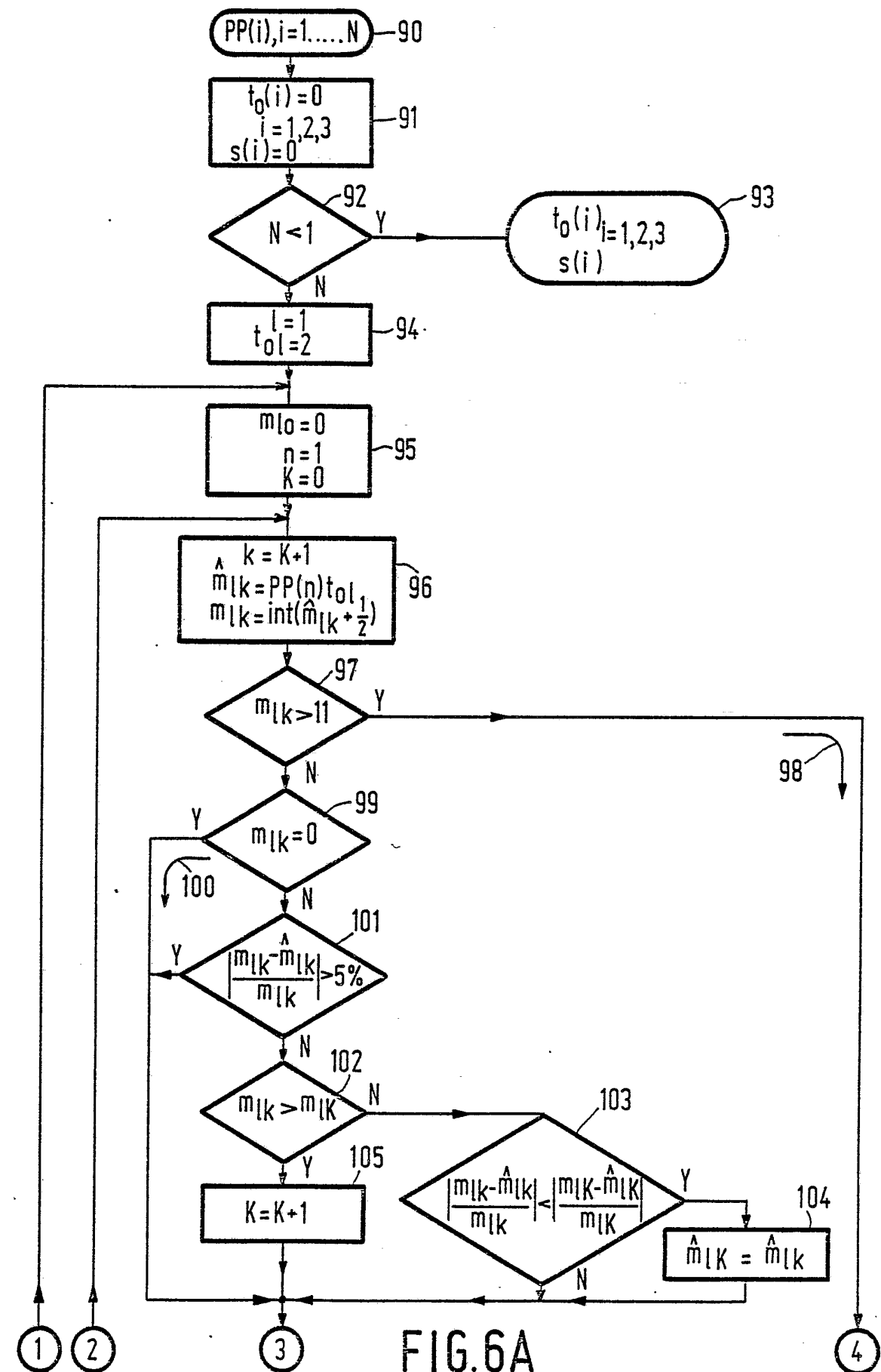


FIG.4B





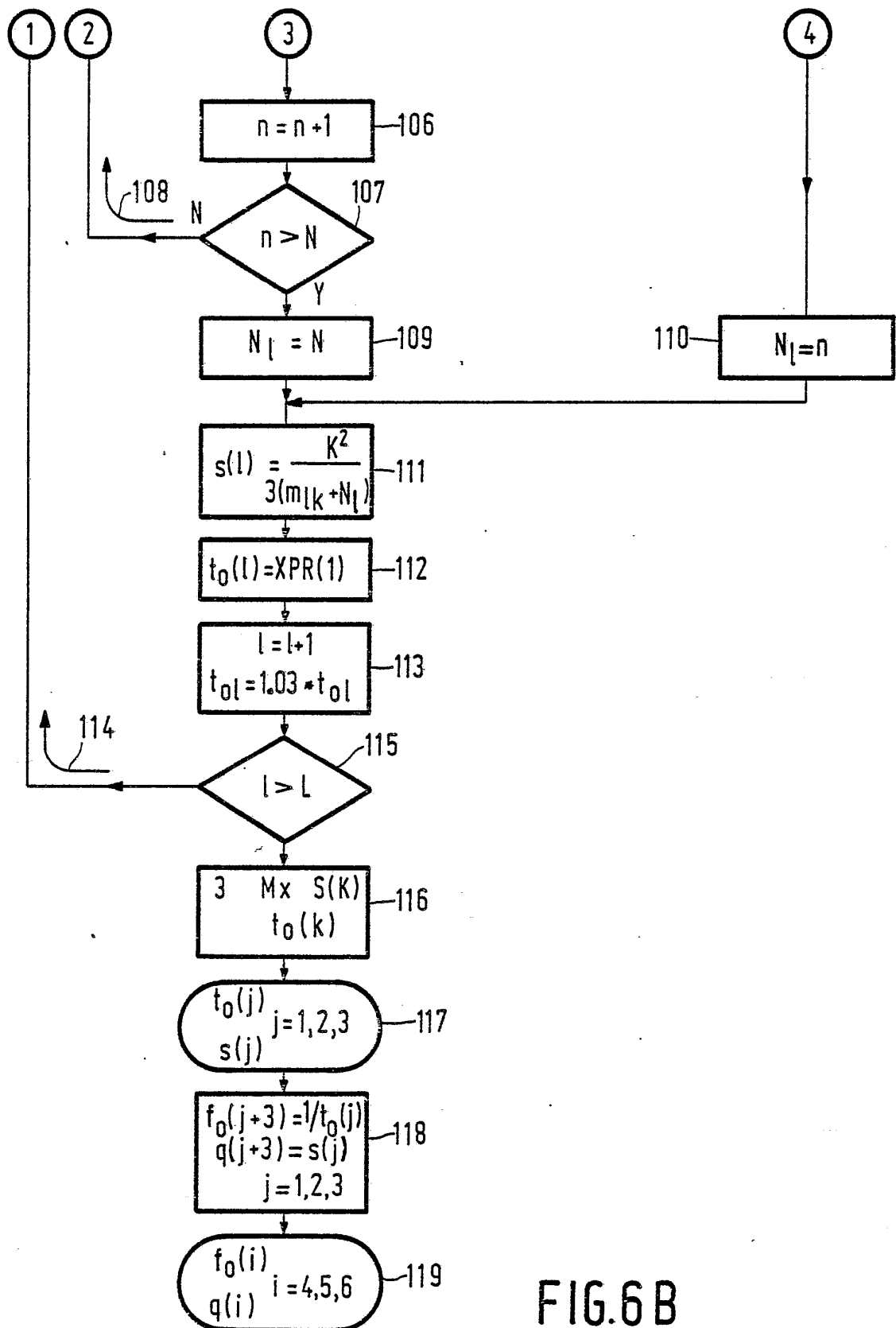


FIG. 6B

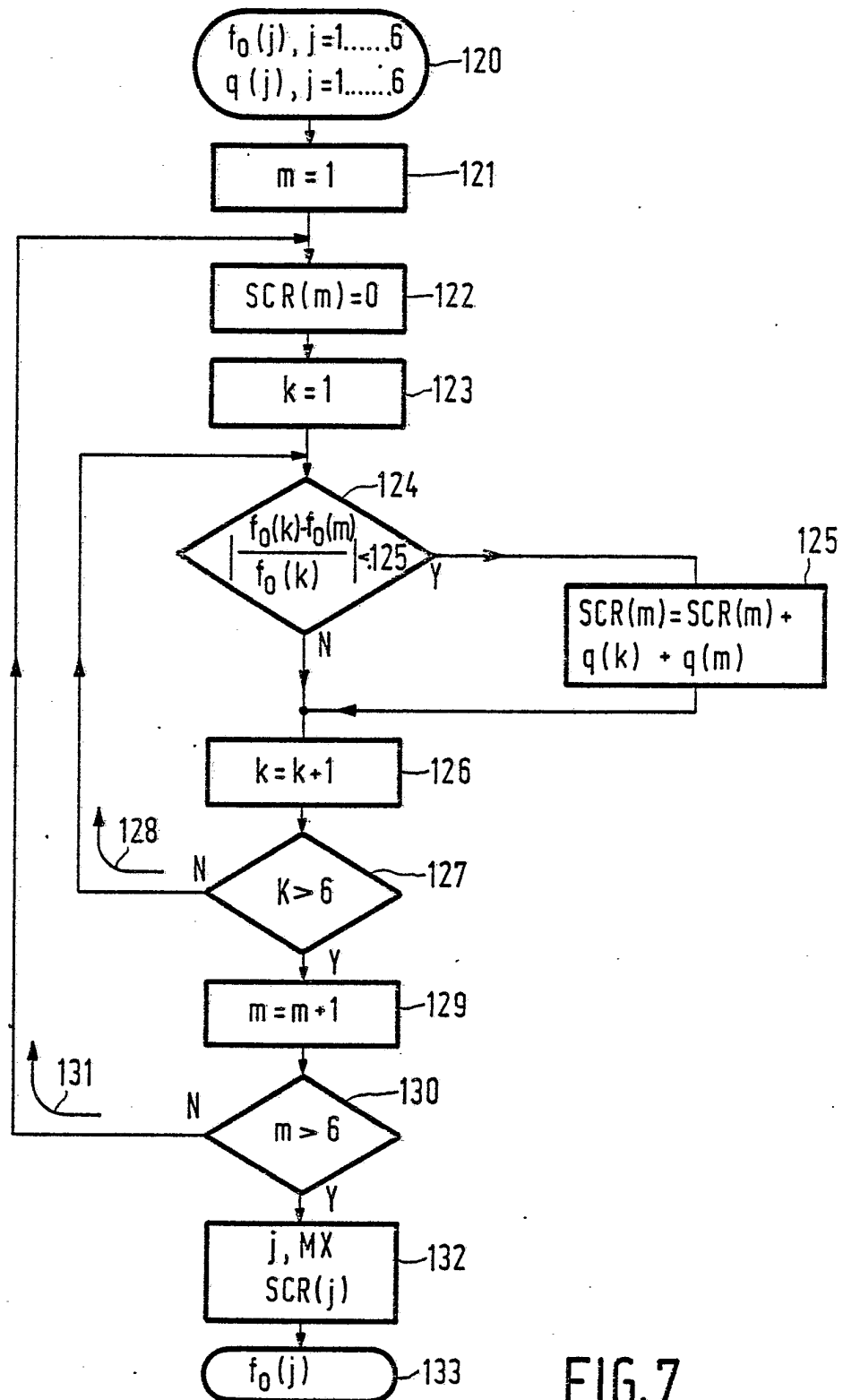


FIG. 7