**Octrooiraad**

(10)A **Terinzagelegging** (11) **8300718**

**Nederland** (19) **NL**

(54) **Werkwijze en inrichting voor herkenning van een foneem in een stemsignaal.**

(51) Int.Cl³.: G10L 1/04, G10L 1/08.

(71) Aanvrager: Sony Corporation (Sony Kabushiki Kaisha) te Tokio.

(74) Gem.: Ir. R. Hoijtink c.s.
Octrooibureau Arnold & Siedsma
Sweelinckplein 1
2517 GK 's-Gravenhage.

(21) Aanvrage Nr. 8300718.

(22) Ingediend 25 februari 1983.

(32) Voorrang vanaf 25 februari 1982.

(33) Land van voorrang: Japan (JP).

(31) Nummer van de voorrangsaanvrage: 29471/82 .

(62) - -

(43) Ter inzage gelegd 16 september 1983.

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

# Werkwijze en inrichting voor herkenning van een foneem in een stemsignaal.

De uitvinding heeft betrekking op een werkwijze en een inrichting voor herkenning van een foneem in een stemsignaal, en meer in het bijzonder op een dergelijke werkwijze en inrichting, welke geschikt zijn voor herkenning met gemak en stelligheid van een foneem in een van een ongeïdentificeerde spreker afkomstig stemsignaal.

Men kent reeds een inrichting voor herkenning van een foneem in een stem- of woordsignaal, dat van een geïdentificeerde spreker afkomstig is. Bij de toepassing van een dergelijke bekende inrichting spreekt een geïdentificeerde spreker alle te herkennen woorden uit, waarbij bepaalde acoustische parameters van de desbetreffende woorden met behulp van een bandfilterbank en dergelijke worden gedetecteerd en opgeslagen, respectievelijk geregistreerd. Wanneer de desbetreffende spreker vervolgens de desbetreffende woorden opnieuw uitspreekt, worden de desbetreffende acoustische parameters van die woorden opnieuw gedetecteerd en vervolgens vergeleken met de eerder opgeslagen of geregistreerde parameters; in geval van coïncidentie wordt een foneem herkend als het door de desbetreffende, geïdentificeerde spreker uitgesproken woord.

Bij een dergelijke inrichting van bekend type worden, wanneer de tijdbasis van een van de spreker afkomstige uiting niet samenvalt met die van de eerder geregistreerde uiting, de uit de met constate tijdsintervallen (5-20 milliseconden) geëxtraheerde, acoustische parameterwaarden bestaande tijdsreeksen aan zodanige compressie of expansie onderworpen, dat de tijdbases komen samen te vallen; op die wijze worden schommelingen in de spreeksnelheid opgevangen.

Bij de bekende inrichting dienen alle in aanmerking komende acoustische parameterwaarden van alle in de herkenning te betrekken woorden vooraf te worden vastgelegd, respectievelijk opgeslagen, zodat een zeer grote geheugencapaciteit vereist is en een betrekkelijk groot aantal mathe-

matische berekeningen dient te kunnen worden uitgevoerd. De bekende inrichting heeft voorts het nadeel, dat in geval van de genoemde tijdbasisaanpassing de uit de met constante tijdsintervallen geëxtraheerde, acoustische parameterwaarden bestaande tijdsreeksen aan de genoemde compressie of expansie dienen te worden onderworpen, hetgeen verdere mathematische berekeningen vereist. Bij onvoldoende overeenstemming tussen de tijdbases bestaat uiteraard het risico van herkenningsfouten.

Naast de in het voorgaande beschreven herkenningsinrichting heeft men bovendien een foneemherkenningsmethode voorgesteld, welke instaat is tot herkenning van individuele fonemen (in de Japanse taal de klinkers A,I,U,E, O, en de medeklinkers K,S,T enz.) of van individuele lettergrepen (KA,KI,KU, enz.)

Deze herkenningsmethoden heeft echter het nadeel, dat klinkerfonemen en dergelijke met quasi-stationaire gedeelten weliswaar betrekkelijk gemakkelijk kunnen worden herkend, doch dat fonemen van betrekkelijk korte duur, zoals voor de plosieve geluiden K,T,P enz., uiterst moeilijk op basis van de acoustische parameterwaarden van slechts één foneem kunnen worden geïdentificeerd.

In verband daarmede is reeds voorgesteld, dat per lettergreep discreet uitgesproken foneem geluiden worden geregistreerd, doch de herkenning van diffuus geuite foneemgeluiden plaatsvindt door soortgelijke tijdbasisaanpassing als reeds beschreven; daarbij wordt de herkenning van de fonemen van korte geluiden, zoals de genoemde plosieve geluiden K,T,P, enz. mogelijk.

De speciale wijze, waarop dergelijke geluiden worden uitgesproken leidt echter tot een beperking van de mogelijkheden van deze methode; daarnaast vereist de genoemde tijdbasisaanspassing steeds een betrekkelijk groot aantal mathematische berekeningen.

Voorts kan omtrent de zojuist beschreven methode worden opgemerkt, dat wanneer deze herkenningsmethode wordt toegepast op de door ongeïdentificeerde sprekers uit-

gesproken woorden, uit de verschillen tussen de betrokenen een zodanige spreiding in acoustische parameterwaarden resulteert, dat geen foneem herkenning uitsluitend door middel van de genoemde tijdbasisaanpassing mogelijk is.

In verband hiermede zijn reeds verschillende verbeteringsvoorstellen gedaan. Volgens een dergelijk voorstel wordt een aantal acoustische parameterwaarden van bijvoorbeeld één woord opgenomen en vindt de herkenning van een foneem plaats op basis van benaderde parameterwaarden; volgens een ander voorstel wordt het gehele woord omgezet in parameterwaarden van vaste grootte en vervolgens geëvalueerd, respectievelijk onderzocht, met behulp van een discriminerende functie. Bij de methoden volgens deze verbeteringsvoorstellen doet zich echter steeds het probleem voor, dat de toepassing van een betrekkelijk grote opslagcapaciteit vereist is en een betrekkelijk groot aantal mathematische berekeningen dient te worden uitgevoerd, terwijl het aantal in de herkenning te betrekken woorden zeer gering is.

De onderhavige uitvinding stelt zich ten doel, hierin verbetering te brengen en een foneemherkenningswijze en -inrichting te verschaffen, welke vrij van de hiervoor genoemde nadelen en beperkingen zijn.

Voorts stelt de uitvinding zich ten doel, een foneemherkenningswijze en -inrichting te verschaffen, waarmede gemakkelijke en stellige herkenning van een foneem mogelijk is zonder op tijdbasisaanpassing gerichte compressie of expansie vanuit acoustische parameterwaarden bestaande tijdsreeksen en zonder de noodzaak van een speciale uitspraak van woorden.

Een ander doel van de uitvinding is het verschaffen van een foneemherkenningswijze en -inrichting, waarmede een gemakkelijke en stellige foneemherkenning mogelijk is, zelfs wanneer deze herkenning op de uiting van een ongeidentificeerde spreker betrekking heeft; daarbij is de genoemde tijdsreekscompressie-expansie overbodig, zodat evenmin een zeer grote opslagcapaciteit voor opslag van acoustische

parameterwaarden vereist is en geen beperking aan het aantal in de herkenning te betrekken woorden behoeft te worden gesteld.

Uitgaande van een werkwijze voor herkenning van een foneem in een stemsignaal onder vorming van een het stemsignaal weergevend electrisch signaal, schrijft de uitvinding nu voor, dat een dergelijke herkenningswijze voorts dient te zijn gekenmerkt door:

extractie uit het electrische signaal van een eerste acoustisch parametersignaal, dat een foneeminformatie van het stemsignaal vertegenwoordigt,

detectie van een stilte-foneem-overgang of een foneem-foneem-overgang in het eerste acoustische parametersignaal,

opwekking van een indicatiesignaal, dat het optreden van een dergelijke overgang aanwijst,

opslag van het eerste acoustische parametersignaal, en door

extractie, op basis van het indicatiesignaal, uit het opgeslagen eerste acoustische parametersignaal van een tweede acoustisch parametersignaal, dat tenminste de stilte-foneem-overgang of de foneem-foneem-overgang van het eerste acoustische parametersignaal bevat.

Voorts verschaft de uitvinding een inrichting voor herkenning van een foneem in een stemsignaal. Uitgaande van een dergelijke inrichting, welke is uitgerust met middelen voor vorming van een het stemsignaal weergevend, electrisch signaal, schrijft de uitvinding voor dat een dergelijke inrichting voorts dient te zijn gekenmerkt door:

eerste extractiemiddelen voor extractie uit het electrische signaal van een eerste acoustisch parametersignaal, dat een foneem informatie van het stemsignaal vertegenwoordigt,

detectiemiddelen voor detectie van een stilte-foneem-overgang of een foneem-foneem-overgang in het eerste acoustische parametersignaal,

opwekmiddelen voor opwekking van een indicatiesignaal, dat het optreden van een dergelijke overgang aanwijst,

opslagmiddelen voor opslag van het eerste acoustische parametersignaal, en door

extractiemiddelen voor extractie, op basis van het indicatiesignaal, uit het opgeslagen eerste acoustische parametersignaal van een tweede acoustisch parametersignaal, dat tenminste de stilte-foneem-overgang of de foneem-foneem-overgang van het eerste acoustische parametersignaal bevat.

De uitvinding zal worden verduidelijkt in de nu volgende beschrijving aan de hand van de bijbehorende tekening. Daarin tonen:

Fig. 1A en 1B schematische weergaven van foneemveranderingen ter verduidelijking van de foneemherkenningswijze volgens de uitvinding,

Fig. 2 een blokschema van een uitvoeringsvorm van een foneemherkenningsinrichting volgens de uitvinding,

Fig. 3A-3H enige schematische weergaven van het ontstaan van acoustische parameterwaarden ter verduidelijking van de werking van een foneemherkenningsinrichting volgens de uitvinding,

Fig. 4 een tabel ter verduidelijking van de werking van een foneemherkenningswijze volgens de uitvinding,

Fig. 5A-5I enige grafieken ter verduidelijking van een foneemherkenningswijze volgens de uitvinding,

Fig. 6 een principeblokschema van een foneemovergangsdetectieschakeling ten behoeve van een foneemherkenningsinrichting volgens de uitvinding en

Fig. 7A-7C enige grafieken van de relatie tussen een foneem en een gedetecteerde parameterwaarde ter verduidelijking van de foneemherkenningswijze volgens de uitvinding.

Voorafgaande aan een meer gedetaileerde beschrijving van de uitvinding wordt eerst ingegaan op de wijze, waarop geluiden (tijdens het spreken) worden geuit.

In de eerste plaats kan worden opgemerkt, dat een geluid kan worden geuit, respectievelijk uitgesproken met grote nadruk op de afzonderlijke klinkers en medeklinkers (S,H, enz.). Zo kan bijvoorbeeld bij de uitspraak van het

woord "HAI" het geluid op de in Fig. 1A schematisch weergegeven wijze variëren volgens "stilte → H → A → I → stilte". In de tweede plaats kan het geluid bij de uitspraak van hetzelfde woord "HAI" op de in Fig. 1B schematisch weergegeven wijze variëren. Hieruit komt naar voren, dat een quasi-stationair deel of segment, bestaande uit foneemgeluiden zoals H,A,I e.d., van uitspraak tot uitspraak in lengte (tijdsduur kan variëren), terwijl een stilte-foneem-overgang of een foneem-foneem-overgang, dat wil zeggen het tussen de quasi-stationaire delen of segmenten in de Fig. 1A en 1B met een schuine lijn gemarkeerde deel of segment, een lengte of tijdsduur vertoont, welke van uitspraak tot uitspraak niet zeer sterk variëert.

Het voorgaande wil zeggen, dat bij uiting van een geluid wel een tijdbasisvariatie van de quasi-stationaire delen of segmenten optreedt, doch niet of in veel geringere mate in de stilte-foneem- of foneem-foneem-overgangen.

De uitvinding is op het zojuist gesignaleerde verschil gebaseerd, waartoe nu eerst naar het blokschema volgens Fig. 2 wordt verwezen.

In Fig. 2 bevat een met een volle lijn getekend blok A een microfoon 1 en een daarop volgende microfoonversterker 2; het desbetreffende blok zet een stemgeluid in een electrisch signaal om. Een met een volle lijn getekend blok B bevat een laagdoorlaatfilter 3, een analoog/digitaal-omzetter 4, een register 6 een snelle-Fourier-transformatieschakeling 8 (FFT) en een detector 9; het blok B abstraheert uit het genoemde electrische signaal een eerste acoustisch parametersignaal. Een met een volle lijn getekend blok C bevat een accentuëringsschakeling 10 en een foneemovergangsdetectieschakeling 20; het blok C dient voor detectie van een stilte-foneem-overgang of een foneem-foneem-overgang in het eerste acoustische parametersignaal. Een met een gebroken lijn getekend blok D bevat eveneens de genoemde accentuëringsschakeling 10, een frequentiebanddeelschakeling 11, een logaritmische schakeling 12, een discrete-Fourier-transforma-

tieschakeling 13 (DFT) en een geheugenschakeling 14; het blok D dient voor detectie van een tweede acoustisch parametersignaal in het eerste acoustische parametersignaal op basis van een door de detectieschakeling 20 afgegeven signaal.

Een van de microfoon 1 afkomstig stemsignaal wordt via de microfoonversterker 2 en het laagdoorlaatfilter 3 tot een frequentiewaarde van minder dan 5,5Khz. doorgelaten naar de analoog/digitaal-omzetter 4, welke van een klokimpulsgenerator 5 een bemonsterklokimpuls met een impulsherhalingsfrequentie van 12,5Khz. en een verschijningsinterval van 80 $\mu$ seconden krijgt toegevoerd; daardoor wordt het stemsignaal in het ritme van de bemonsterklokimpuls omgezet in een digitaal signaal met een voorafbepaald aantal bits per woord. Het aldus aan omzetting onderworpen stemsignaal wordt toegevoerd aan een schuifregister 6 met een capaciteit van 5 x 64 woorden; door de klokimpulsgenerator 5 wordt bovendien een frameklokimpuls met een verschijningsinterval van 5,12 milliseconden aan een telkens-vijf-teller 7 toegevoerd, waarvan het teluitgangssignaal aan het register 6 wordt toegevoerd, zodanig, dat daardoor het stemsignaal met 64 woorden per keer wordt verschoven, zodat het register 6 een verschoven stemsignaal van 4 x 64 woorden afgeeft.

Dit verschoven stemsignaal van 4 x 64 = 256 woorden wordt toegevoerd aan de snelle-Fourier-transformatieschakeling 8 (FFT). Indien nu wordt aangenomen, dat een uit kleine $n_f$ monsterinformatiewaarden bestaande golfvormfunctie U, welke zich over een tijdsduur T uitstrekt, kan worden weergegeven als:

$$U_{nf}T(f) \qquad (1),$$

dan leidt Fourier-transformatie van de golfvormfunctie tot een signaal, dat kan worden weergegeven als:

$$U_{nf}T(f) = \int_{-T/2}^{+T/2} U_{nf}T(f)e^{-2\pi jft}dt$$

$$= U_{1nf}T(f) + jU_{2nf}T(f) \qquad (2).$$

Het uitgangssignaal van de snelle-Fourier-transfor-

matieschakeling 8 wordt toegevoerd aan de energiespectrum-signaaldetectieschakeling 9, waarvan het uitgangssignaal een energiespectrumsignaal vormt, waarvoor geldt:

$$|U^2| = U^2_{1nf}T(f) + U^2_{2nf}T(f) \qquad (3)$$

Aangezien het uit deze Fourier-transformatie resulterende signaal symmetrisch ten opzichte van de frequentie-as is, is de helft van de $n_f$ uit de transformatie resulterende monsterwaarden redundant; uitsluiting van de helft van de $n_f$ monsterwaarden resulteert dan in de levering van $1/2n_f$ informatiewaarden. Het 256-woordssignaal, dat aan de genoemde snelle-Fourier-transformatieschakeling 8 wordt toegevoerd, resulteert derhalve na de transformatie in een 128-woords energiespectrumsignaal.

Dit energiespectrumsignaal wordt toegevoerd aan de accentueringsschakeling 10, welke een zodanige weging van het signaal uitvoert, dat correctie voor "auditory sens" wordt verkregen. Als voorbeeld van een dergelijke weging kan een correctie worden genoemd, waarbij bijvoorbeeld de hoogfrequentcomponent van het signaal wordt geäccentueerd.

Het aldus aan weging onderworpen signaal wordt toegevoerd aan de frequentiebanddeelschakeling 11, welke het signaal bijvoorbeeld verdeelt over 32 frequentiebanden volgens een voor geluidswaarneming geschikte frequentie "mel-scale". Wanneer deze frequentiebanden niet samenvallen met de deelpunten van het energiespectrum, wordt het signaal in zodanige frequentiebanden opgesplitst, dat met de verdeling van het signaal over de respectieve frequentiebanden overeenkomende signalen worden verkregen, zodanig, dat het oorspronkelijke 128-woords energiespectrumsignaal wordt gecomprimeerd tot een energiespectrumsignaal van 32 woorden met acoustische eigenschappen.

Dit laatstgenoemde signaal wordt toegevoerd aan de logaritmische schakeling 12 voor omzetting van ieder signaal in logaritmische waarden. De door de weging en dergelijke in de accentueringsschakeling 10 veroorzaakte redundantie van het energiespectrumsignaal wordt derhalve uitgesloten bij

weergave van het gelogaritmiseerde energiespectrum

$$\log \left| U^2_{nf} T(f) \right| \qquad (4)$$

door de spectrumparameter $x_{(i)}$, waarbij i = 0,1,...,31.

Deze spectrumparameter $x_{(i)}$ wordt toegevoerd aan de discrete-Fourier-transformatieschakeling 13 (DFT). Indien daarbij het aantal uit de verdeling resulterende frequentiebanden M bedraagt, voert de discrete-Fourier-transformatieschakeling 13 een discrete Fourier-analyse van 2M-2 punten uit, waarbij de M-dimensionale parameter $x_{(i)}$ (i = 0,1, ...,M-1) als het reëele aantal in 2M-1 punten symmetrische parameterwaarden geldt. Dit wil zeggen:

$$X_{(m)} = \sum_{i=0}^{2M-3} x_{(i)} W^{mi}_{2M-2} \qquad (5),$$

waarin $W^{mi}_{2M-2} = e^{-j(\frac{2\pi . i.m}{2M-2})}$, waarbij m = 0,1,...,2M-3.

Aangezien de functie,volgens welke de discrete Fourier-transformatie wordt uitgevoerd, als een even functie wordt beschouwd, leidt het voorgaande tot:

$$W^{mi}_{2M-2} = \cos(\frac{2\pi . i.m}{2M-2}) = \cos \frac{\pi . i.m}{M-1} ,$$

hetgeen leidt tot:

$$X_{(m)} = \sum_{i=0}^{2M-3} x_{(i)} \cos \frac{\pi . i.m}{M-1} \qquad (6).$$

Door deze discrete Fourier-transformatie (DFT) worden de acoustische parameterwaarden geëxtraheerd, welke de omhullende van het spectrum karakteriseren.

Voor de spectrumparameter $x_{(i)}$, welke op de hier beschreven wijze aan discrete Fourier-transformatie onderworpen wordt, worden de waarden voor de P dimensies van 0 tot P-1 (bijvoorbeeld P=8), geëxtraheerd en samengesteld tot de locale parameter $L_{(p)}$ (p = 0,1,...,P-1) van de gedaante:

$$L_{(p)} = \sum_{i=0}^{2M-3} x_{(i)} \cos \frac{\pi . i . p}{M-1} \qquad (7).$$

Het feit, dat de spectrumparameter symmetrisch is, leidt tot:

$x_{(i)} = x_{(2M-i-2)}$ (8), hetgeen tot een verandering van de locale parameterwaarden $L_{(p)}$ leidt tot:

$$L_{(p)} = X(\phi) + \sum_{i=1}^{M-2} x_{(i)} \{\cos \frac{\pi \cdot i \cdot p}{M-1} + \cos \frac{\pi (2M-2-i) p}{M-1}\} + X(M-1) \cos \frac{\pi \cdot p}{M-1},$$

waarin p = 0,1,....,P-1.

Op deze wijze heeft compressie van het 32-woords signaal tot een P-woords signaal, bijvoorbeeld een 8-woords signaal, plaatsgevonden.

De desbetreffende locale parameterwaarden $L_{(p)}$ worden toegevoerd aan de geheugenschakeling 14. Deze bevat een matrixverdeling van geheugensecties met bijvoorbeeld 16 rijen, welke elk uit P-woorden bestaan, waarin de locale parameterwaarden $L_{(p)}$ voor iedere dimensie om de beurt worden opgeslagen; de frameklokimpuls met een verschijningsinterval van 5,12m seconden wordt door de genoemde klokimpulsgenerator 5 geleverd, zodat de parameterwaarden van iedere rij in zijdelingse richting worden verplaatst. In de geheugenschakeling 14 vindt derhalve opslag plaats van de locale parameterwaarden $L_{(p)}$ voor P dimensies met een interval van 5,12m sec. dit geschiedt in de vorm van 16 frames (81,92m sec.). De desbetreffende locale parameterwaarden $L_{(p)}$ worden bij het verschijnen van iedere volgende frameklokimpuls "bijgewerkt" (updated).

Het bijvoorbeeld van de accentueringsschakeling 10 afkomstige signaal wordt bovendien toegevoerd aan de foneemovergangsdetectieschakeling 20 voor detectie van de overgang tussen opeenvolgende fonemen.

Het uitgangssignaal van de schakeling 20, respectievelijk het overgangsdetectiesignaal $T_{(t)}$, wordt toe-

gevoerd aan de geheugenschakeling 14, zodanig, dat op het tijdstip, waarop de bij het verschijnen van dit detectiesignaal behorende locale parameterwaarden $L_{(p)}$ naar de achtste rij wordt doorgeschoven, uitlezing van de geheugenschakeling 14 plaatsvindt. Een dergelijke uitlezing van de geheugenschakeling 14 heeft de gedaante van de uitlezing van 16 frames in zijdelingse richting voor iedere dimensie P; de aldus uitgelezen signalen worden toegevoerd aan de discrete-Fourier-transformatieschakeling 15 (DFT).

Deze schakeling 15 voert op soortgelijke wijze discrete Fourier-transformatie uit, zodat de omhullende van de tijdsreeksverandering van de acoustische parameterwaarden wordt verkregen. Uit de desbetreffende DFT-signalen worden de waarden voor Q dimensies van 0 tot Q-1 verkregen, waarbij bijvoorbeeld Q=3. Deze digitale Fourier-transformatie vindt voor iedere dimensie P plaats, waaruit overgangsparameterwaarden $K_{(p,q)}$ resulteren (p=0,1,..,P-1 en q=0,1,...,Q-1) voor in totaal P x Q (=24) woorden. Daarbij kunnen, aangezien $K_{(0,0)}$ de macht van de stemgolfvorm vertegenwoordigt, ter wille van energienormalisering voor $p_0$ de waarden q=1 tot Q worden verkregen.

Onder verwijzing naar de schematische weergave volgens de Fig. 3A-3H wordt opgemerkt, dat dit wil zeggen, dat wanneer de overgang volgens Fig. 3B van een ingangsstemsignaal (HAI) volgens Fig. 3A wordt gedetecteerd, het totale energiespectrum van dit signaal bijvoorbeeld de gedaante volgens Fig. 3C heeft. Indien het energiespectrum van de overgang van "H⟶A" de gedaante volgens Fig. 3D heeft, krijgt het desbetreffende signaal na accentuering de gedaante volgens Fig. 3E; na compressie volgens de "mel-scale" resulteert de gedaante volgens Fig. 3F. Het desbetreffende signaal krijgt na discrete Fourier-transformatie de gedaante volgens Fig. 3G. De 16 voor- en achterframes van dit signaal hebben na matrixbewerking de gedaante volgens Fig. 3H, waarna discrete Fourier-transformatie in de richting van de tijdbasis of as t tot de overgangsparameterwaarden $K_{(p,q)}$ leidt.

Deze overgangsparameterwaarden $K_{(p,q)}$ worden toegevoerd aan een berekeningsschakeling 16 voor berekening van de afstand volgens Mahalanobis; de berekeningsschakeling 16 krijgt bovendien van een geheugeninrichting 17 een "cluster coëfficiënt" toegevoerd voor berekening van de genoemde afstand volgens Mahalanobis voor ieder van deze coefficiënten; bij een dergelijke berekening resulteert de desbetreffende coëfficiënt uit aftrekking van de overgangsparameterwaarden van de uitingen van verschillende sprekers, klassifikatie van de overgangsparameterwaarden op basis van het foneembestand en daarop volgende statistische analyse daarvan.

De berekende afstand volgens Mahalanobis wordt toegevoerd aan een evaluatieschakeling 18, waardoor wordt onderzocht of een gedetecteerde overgang een fenomeen-fenomeen-overgang is; het detectie-uitgangssignaal komt ter beschikking aan een uitgangsaansluiting 19.

Meer in het bijzonder worden voor bijvoorbeeld de 12 woorden "HAI", "IIE" en "0(ZERO)"-"9(KYU)" de stemsignalen van een aantal (meer dan honderd) sprekers vooraf aan de beschreven inrichting toegevoerd, waarbij de optredende overgangen worden gedetecteerd en de desbetreffende overgangsparameterwaarden worden geëxtraheerd. Deze overgangsparameterwaarden worden volgens een tabel, bijvoorbeeld de tabel volgens Fig. 4, geklassificeerd en vervolgens voor iedere desbetreffende klassificatie (cluster) aan statistische analyse onderworpen. In de tabel volgens Fig. 4 heeft het symbool ✠ betrekking op stilte.

Voor de desbetreffende overgangsparameterwaarden wordt als arbitrair monster $R_{r,n}^{(a)}$ (r=1,2,..,24 en a vertegenwoordigt de clusterindex; a=1 komt bijvoorbeeld overeen met ✠→H, a=2 komt overeen met H A en n vertegenwoordigt het aantal sprekers) de covariantiematrix

$$A_{r,s}^{(a)} = E(R_{r,n}^{(a)} - \overline{R_r^{(a)}})(R_{s,n}^{(a)} - \overline{R_s^{(a)}}) \qquad (15)$$

berekent, waarin $\overline{R_r^{(a)}} = E(R_{r,n}^{(a)})$ en E een ensemble-gemiddelde

vertegenwoordigt. Vervolgens wordt de inverse-matrix

$$B^{(a)}_{r,s} = (A^{(a)}_{t,n})^{-1}_{r,s} \quad \text{gezocht.} \qquad (16)$$

Op deze wijze wordt de afstand tussen een willekeurige overgangsparameterwaarde $K_r$ en een "cluster" a verkregen als een Mahalanobis-afstand met de gedaante:

$$D(K_r,a) \stackrel{d}{=} \sum_r \sum_s (K_r - \overline{R^{(a)}_r}).B^{(a)}_{r,s} . (K_r - \overline{R^{(a)}_s}) \qquad (17)$$

Indien de genoemde $B^{(a)}_{r,s}$ en $\overline{R^{(a)}_r}$ worden gezocht en vervolgens in de geheugeninrichting 17 worden opgeslagen, wordt de Mahalanobis-afstand tussen de willekeurige overgangsparameterwaarden van het ingangsstemsignaal en het "cluster" berekend door de berekeningsschakeling 16.

Deze laatstgenoemde schakeling verschaft derhalve voor ieder ontvangen stemsignaal de minimale afstand van iedere overgangswaarde tot ieder "cluster" en voorts de volgorde van de overgangen; deze informaties worden vervolgens toegevoerd aan de evaluatieschakeling 18 voor herkenning en beoordeling wanneer het ontvangen stemsignaal stilvalt. Bijvoorbeeld wordt bij ieder woord de afstand berekent als de gemiddelde waarde van de vierkantswortel van de minimale afstand tussen de desbetreffende overgangsparameterwaarde en de clusters. Voor het geval, dat de overgangen gedeeltelijk wegvallen, wordt de woordafstand van het woord onderzocht voor een aantal mogelijk vervallen types. Daarbij wordt echter een woord met een van die volgens de tabel afwijkende volgorde van overgangswaarden afgewezen. Vervolgens wordt het woord met de minimale woordafstand herkend en onderzocht.

Bij een inrichting volgens de uitvinding zal derhalve, aangezien detectie van foneemverandering aan de overgangen van het foneem wordt toegepast, nimmer een tijdbasisschommeling optreden, zodat de van ongeïdentificeerde sprekers afkomstige fonemen op bevredigende wijze kunnen worden herkend.

Aangezien de parameterwaarden op de beschreven wijze bij de overgangen worden geëxtraheerd, kan iedere over-

gang in 24 dimensies worden herkend, met als gevolg, dat de herkenning gemakkelijk en met hoge nauwkeurigheid geschiedt.

Met behulp van de hiervoor beschreven inrichting volgens de uitvinding werd bij een beproeving, waaraan in eerste instantie 120 sprekers deelnamen en vervolgens andere sprekers dan deze 120 aan een onderzoek met 120 woorden werden onderworpen, een gemiddeld herkenningspercentage van 98,2% bereikt.

Bij het hiervoor beschreven voorbeeld kunnen de overgangen "H→A" in "HAI" en "H→A" in "8(HACHI)" beide als tot hetzelfde "cluster" behorend worden geklassificeerd. Indien nu het aantal fonemen van te herkennen woorden $\propto$ bedraagt en vooraf "clusters" van ongeveer $\propto P_2$ fonemen vooraf worden berekend, waarna de daarbij gevonden clustercoëfficiënt in de geheugeninrichting 17 wordt opgeslagen, is dit voldoende voor de herkenning van verschillende woorden, welke vervolgens zonder problemen kunnen worden herkend.

Fig. 6 toont een principeblokschema van een foneemovergangsdetectieschakeling 20, welke bij de foneemherkenningsinrichting volgens de uitvinding kan worden toegepast.

Voorafgaande aan de beschrijving van een dergelijke foneemovergangsdetectieschakeling 20 wordt eerst opgemerkt, dat volgens een gebruikelijke methode van overgangsdetectie gebruik gemaakt wordt van de som van de veranderingshoeveelheden van bijvoorbeeld de acoustische parameterwaarden $L_{(p)}$. Dit wil zeggen, dat wanneer voor ieder frame de parameterwaarden voor P dimensies worden afgetrokken, waarbij de parameter voor het frame G de waarde $L_{(p)}(G)$, waarbij $p=0,1,\ldots,P-1$, de overgangsdetectie plaatsvindt op basis van de som van de absolute waarden van de verschilhoeveelheden, welke som wordt bepaald door:

$$T(G) = \sum_{p-0}^{p-1} \left| L_{(p)}\ (G) - L_{(p)}\ (G-1) \right| \qquad (9')$$

Wanneer P=1, dat wil zeggen één dimensie betreft, zoals de Fig. 5A en 5B laten zien, worden de piekwaarden van

de parameter T(G) verkregen in de punten, waarin de parameterwaarden $L_{(p)}(G)$ verandert. Wanneer P=2, respectievelijk sprake van twee dimensies is, zal de parameterwaarde T(G), indien ondanks soortgelijke verandering als hiervoor van de parameterwaarden $L_{(0)}(G)$ en $L_{(1)}(G)$ voor 0 en 1 volgens respectievelijk de Fig. 5C en 5D optreedt, de verschilhoeveelheden volgens de respectieve Fig. 5E en 5F veranderen, twee pieken vertonen, zodat geen overgang voor één punt kan worden bepaald. Dit verschijnsel zal bijvoorbeeld optreden wanneer de parameterwaarden voor meer dan twee dimensies worden genomen.

Hoewel bij een dergelijke beschouwing tewerk gegaan wordt, alsof de parameterwaarden $L_{(p)}(G)$ een continue gedrag vertoont, vertoont de parameterwaarde $L_{(p)}(G)$ in de praktijk een verandering in discrete stappen. Voorts kan algemeen gesteld worden, dat een foneem een betrekkelijk geringe fluctuatie vertoont, zodat de parameterwaarde $L_{(p)}(G)$ in de praktijk verandert als weergegeven in Fig. 5H, hetgeen leidt tot het optreden van een aantal "holten" en "bolten" in de gedetecteerde parameterwaarde T(G), zoals Fig. 5I laat zien.

Als gevolg daarvan vertoont een dergelijke methode de nadelen, dat geen nauwkeurige detectie wordt verkregen en het detectieniveau niet stabiel is.

In verband daarmede bevat een foneemovergangsdetectieschakeling 20 volgens de onderhavige uitvinding enige deelschakelingen, welke een gemakkelijke en stabiele foneemovergangsdetectie mogelijk maken.

Bij de foneemovergangsdetectieschakeling volgens Fig. 6 wordt het van de accentueringsschakeling 10 volgens Fig. 2 afkomstige, gewogen signaal via een ingangsaansluiting 21-a toegevoerd aan een banddeelschakeling 21, waardoor het signaal in N (bijvoorbeeld 20) banden wordt verdeeld volgens de "mel-scale", waaruit een aan de signaal hoeveelheid per respectieve band toegevoegd signaal $V_{(n)}$ resulteert, waarbij n=0,1,...,N-1. Dit signaal $V_{(n)}$ wordt toegevoerd aan een instelspanningslogaritmeschakeling 22 ter verkrijging van een signaal

$$V'_{(n)} = \log (V_{(n)} + B) \qquad (10)$$

Het signaal $V_{(n)}$ wordt tevens toegevoerd aan een accumulator 23 voor vorming van een signaal $V_{(a)}$ van de gedaante:

$$V_a = \sum_{n-1}^{20} V_{(n)}/20$$

Toevoer van dit signaal $V_a$ aan de instelspanningslogaritmeschakeling 22 levert als resultaat:

$$v'_a = \log (V_a + B) \qquad (11)$$

Verdere toevoer van deze signalen aan een bewerkingsschakeling 24 leidt tot:

$$v_{(n)} = v'_a - v'_{(n)} \qquad (12)$$

Het hiervoor beschreven gebruik van het door de banddeelschakeling 21 geleverde signaal $V_{(n)}$ heeft tot gevolg, dat de veranderingshoeveelheid voor iedere dimensie (n=0,1,...N-1) van dit signaal voor de overgang van foneem tot foneem in bij benadering dezelfde mate wordt verminderd, zodanig, dat de door verschillen tussen fonemen veroorzaakte veranderingshoeveelheid geen spreidingsverschijnselen gaat vertonen. Aangezien eerst wordt gelogaritmiseerd en vervolgens de berekening wordt uitgevoerd voor vorming van de genormaliseerde parameterwaarde $v_{(n)}$, kan worden voorkomen, dat deze parameterwaarden $v_{(n)}$ met veranderingen in het niveau van een ontvangen stemsignaal fluctueert. Aangezien de berekening wordt uitgevoerd onder toevoeging van een instelspanningsniveau B, is het mogelijk, zoals duidelijk wordt uit het feit dat voor $B \to \infty$, $v_{(n)} \to 0$, de gevoeligheid voor betrekkelijk zwakke componenten (ruis en dergelijke) van het ontvangen stemsignaal te verminderen.

De parameterwaarden $v_{(n)}$ worden toegevoerd aan een geheugeninrichting 25 met een capaciteit voor opslag van parameterwaarden voor 2w + 1 (bijvoorbeeld 9) frames. Het uit deze opslag resulterende signaal wordt toegevoerd aan een bewerkingsschakeling 26 voor vorming van een signaal:

$$Y_{n,t} = \min_{I \in GF_N} \{v_{(n)}(I)\} \qquad (13),$$

waarin $GF_N = \{I; -w + t \leqq I \leqq w + t\}$

Toevoer van dit uitgangssignaal van de bewerkingsschakeling 26 en het rechtstreeks van de geheugeninrichting 25 afkomstige parametersignaal $Y_{(e)}$ aan een bewerkingsschakeling 27 levert een signaal:

$$T_{(t)} = \sum_{n=0}^{N-1} \sum_{I=-w}^{w} (v_{(n)} (I + t) - Y_{n,t}) \qquad (14)$$

Dit signaal $T_{(t)}$ vormt de overgangsdetectieparameterwaarde en wordt toegevoerd aan een piekwaarde-evaluatieschakeling 28 voor detectie van een foneemovergang in het ingangsstemsignaal, dat vervolgens aan een uitgangsaansluiting 29 ter beschikking komt voor toevoer aan de uitgangsschakeling van de geheugeninrichting 14 volgens Fig. 2.

Aangezien de parameterwaarde $T_{(t)}$ wordt bepaald door w frames, ieder over het frame t, treden geen onnoodzakelijke "holten","bolten" en "multipolen" op. De Fig. 7A-7C verduidelijken het geval, waarin de uitspraak of uiting van bijvoorbeeld het woord "ZERO" wordt opgenomen als 12-bits digitale informatie met een bemonsterfrequentie van 12,5Khz.; de informatie wordt voor 256 punten aan snelle Fourier-transformatie onderworpen met een frameperiodeduur van 5,12m sec., terwijl de beschreven detectie wordt uitgevoerd voor een aantal banden N=20, een instelspanningswaarde B=0 en een aantal gedetecteerde frames van 2w+1=9. Fig. 7A toont de stemgeluidsgolfvormen, Fig. 7B de fonemen en Fig. 7C het gedetecteerde signaal, waarin de opmerkelijke piekwaarden optreden bij de respectieve overgangen "stilte → Z", "Z → E", "E → R", "R → O" en "O → stilte". Hoewel in het stilte-deel als gevolg van ruis enige "holten" en "bolten" voorkomen, kunnen deze practisch tot nagenoeg de waarde nul worden teruggebracht door verhoging van het instelspanningsniveau B, zoals met een gebroken lijn in Fig. 7C is aangeduid.

Het voorgaande beschrijft de detectie van foneemovergangen. Bij de daartoe toegepaste foneemovergangsdetectie schakeling 20 volgens de uitvinding is stabiele detectie van foneemovergangen met geringe schommeling van de detectie-

parameterwaarden als gevolg van verschillen in fonemen en van optredende niveauveranderingen op ieder ogenblik mogelijk.

Bovendien beperkt de foneemovergangsdetectievolgens de uitvinding zich niet tot de hiervoor beschreven foneemherkenningswijze, doch kan een dergelijke detectie eveneens worden toegepast in gevallen, waarin het stationaire segment of deel tussen gedetecteerde overgangen zelf het object van detectie vormt en de tijdbases van de stationaire segmenten aan elkaar worden aangepast door gebruik making van de gedetecteerde overgangen. De foneemovergangsdetectieschakeling volgens de uitvinding kan bovendien met voordeel worden toegepast bij de analyse van overgangen in geval van stemgeluidssynthese.

De uitvinding beperkt zich niet tot de in het voorgaande beschreven en in de tekening weergegeven,enkele uitvoeringsvorm van de uitvinding. Verschillende wijzigingen kunnen in de beschreven details en in de onderlinge samenhang daarvan worden aangebracht, zonder dat daarbij het kader van de uitvinding wordt overschreden.

--

CONCLUSIES

1. Werkwijze voor herkenning van een foneem in een stemsignaal onder vorming van een, het stemsignaal weergevend, electrisch signaal, g e k e n m e r k t door:

extractie uit het electrische signaal van een eerste acoustisch parametersignaal, dat een foneeminformatie van het stemsignaal vertegenwoordigt,

detectie van een stilte-foneem-overgang of een foneem-foneem-overgang in het eerste acoustische parametersignaal,

opwekking van een indicatiesignaal, dat het optreden van een dergelijke overgang aanwijst,

opslag van het eerste acoustische parametersignaal, en door

extractie, op basis van het indicatiesignaal, uit het opgeslagen eerste acoustische parametersignaal van een tweede acoustisch parametersignaal, dat tenminste de stilte-foneem-overgang of de foneem-foneem-overgang van het eerste acoustische parametersignaal bevat.

2. Herkenningswijze volgens conclusie 1, m e t h e t k e n m e r k, dat de extractie van het eerste acoustische parametersignaal plaatsvindt door:

omzetting van het in analoge vorm verkerende, electrische signaal in een digitaal signaal,

opslag van een aantal dergelijke digitale signalen en

vorming van het eerste acoustische parametersignaal door Fourier-transformatie van een aantal opgeslagen digitale signalen.

3. Herkenningswijze volgens conclusie 1, m e t h e t k e n m e r k, dat de detectie van een overgang plaatsvindt door:

afscheiding van een energieniveausignaal voor ieder van een aantal frequentiebanden uit het eerste acoustische parametersignaal,

berekening van het gemiddelde van de energieniveausignalen, gevolgd door berekening van een aantal eerste ver-

schilniveau's tussen het berekende gemiddelde van de energieniveausignalen en die respectieve energieniveausignalen,

extractie van het laagste van de berekende eerste verschilniveau's,

berekening van een aantal tweede verschilniveau's tussen dat laagste niveau en de respectieve eerste verschilniveau's,

vorming van een overgangsdetectieparametersignaal op basis van de tweede verschilniveau's voor de energieniveausignalen van het eerste acoustische parametersignaal, en door

detectie van een stilte-foneem-overgang of een foneem-foneem-overgang op basis van het overgangsdetectieparametersignaal.

4. Herkenningswijze volgens conclusie 3, m e t h e t k e n m e r k, dat de detectie van een overgang geschiedt door accentuering van het energieniveau van het eerste acoustische parametersignaal.

5. Herkenningswijze volgens conclusie 1, m e t h e t k e n m e r k, dat de opslag van het eerste acoustische parametersignaal geschiedt door:

scheiding van het eerste acoustische parametersignaal in een aantal frequentiebandsignalen,

omzetting van de frequentiebandsignalen door Fouriertransformatie in een derde acoustisch parametersignaal, en door

ontvangst van het derde acoustische parametersignaal en opslag van een aantal dergelijke derde acoustische parametersignalen.

6. Herkenningswijze volgens conclusie 5, g e k e n m e r k t door weging van het eerste acoustische parametersignaal.

7. Inrichting voor electrische herkenning van een foneem in een stemsignaal, voorzien van middelen voor vorming van een het stemsignaal weergevend, electrisch signaal, g e k e n m e r k t door:

middelen voor extractie uit het electrische signaal van een eerste acoustische parametersignaal, dat een foneem-

informatie van het stemsignaal vertegenwoordigt,

detectiemiddelen voor detectie van een stilte-foneem-overgang of een foneem-foneem-overgang in het eerste acoustische parametersignaal en voor afgifte van een indicatiesignaal, dat het optreden van een dergelijke overgang aanwijst, en door

opslagmiddelen voor opslag van het eerste acoustische parametersignaal en voor extractie, op basis van het indicatiesignaal, uit het opgeslagen signaal van een tweede acoustisch parametersignaal, dat tenminste de stilte-foneem-overgang of de foneem-foneem-overgang van het eerste acoustische parametersignaal bevat.

8. Herkenningsinrichting volgens conclusie 7, m e t h e t k e n m e r k, dat de extractiemiddelen omvatten:

omzetmiddelen voor omzetting van het in analoge vorm verkerende, electrische signaal in een digitaal signaal,

opslagmiddelen voor opslag van een aantal dergelijke digitale signalen en

extractiemiddelen voor vorming van het eerste acoustische parametersignaal door Fourier-transformatie van de opgeslagen digitale signalen.

9. Herkenningsinrichting volgens conclusie 7, m e t h e t k e n m e r k, dat de detectiemiddelen zijn voorzien van:

middelen voor afscheiding van een energieniveausignaal voor ieder van een aantal frequentiebanden uit het eerste acoustische parametersignaal,

berekeningsmiddelen voor berekening van het gemiddelde van de energieniveausignalen en voor berekening van een aantal eerste verschilniveau's tussen dat gemiddelde en de respectieve energieniveausignalen,

extractiemiddelen voor extractie van het laagste van de eerste verschilniveau's,

berekeningsmiddelen voor berekening van een aantal tweede verschilniveau's tussen het geëxtraheerde laagste niveau en de respectieve eerste verschilniveau's en voor

afgifte van een overgangsdetectieparametersignaal op basis van de tweede verschilniveau's voor de energieniveausignalen van het eerste acoustische parametersignaal, en van

detectiemiddelen voor detectie van een stilte-foneem-overgang of een foneem-foneem-overgang op basis van het overgangsdetectieparametersignaal en voor afgifte van het indicatiesignaal.

10. Herkenningsinrichting volgens conclusie 9, m e t h e t k e n m e r k, dat de detectiemiddelen middelen voor accentuering van het energieniveau van het eerste acoustische parametersignaal omvatten.

11. Herkenningsinrichting volgens conclusie 7, m e t h e t k e n m e r k, dat de opslagmiddelen zijn voorzien van:

scheidingsmiddelen voor scheiding van het eerste acoustische parametersignaal in een aantal energieniveausignalen,

omzetmiddelen voor omzetting van de energieniveausignalen met behulp van Fourier-transformatie in een derde acoustische parametersignaal, en van

opslagmiddelen voor opslag van een aantal dergelijke derde acoustische parametersignalen.

12. Herkenningsinrichting volgens conclusie 11, g e k e n m e r k t door weegmiddelen voor weging van het eerste acoustische parametersignaal.

---

FIG. 1A

FIG. 1B

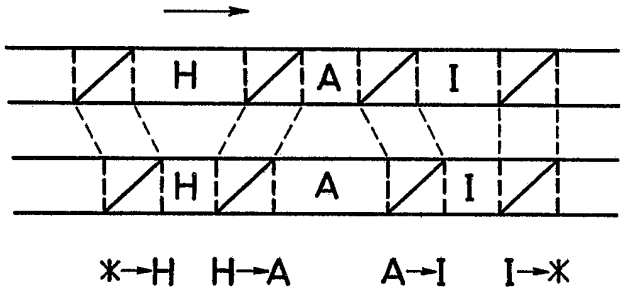
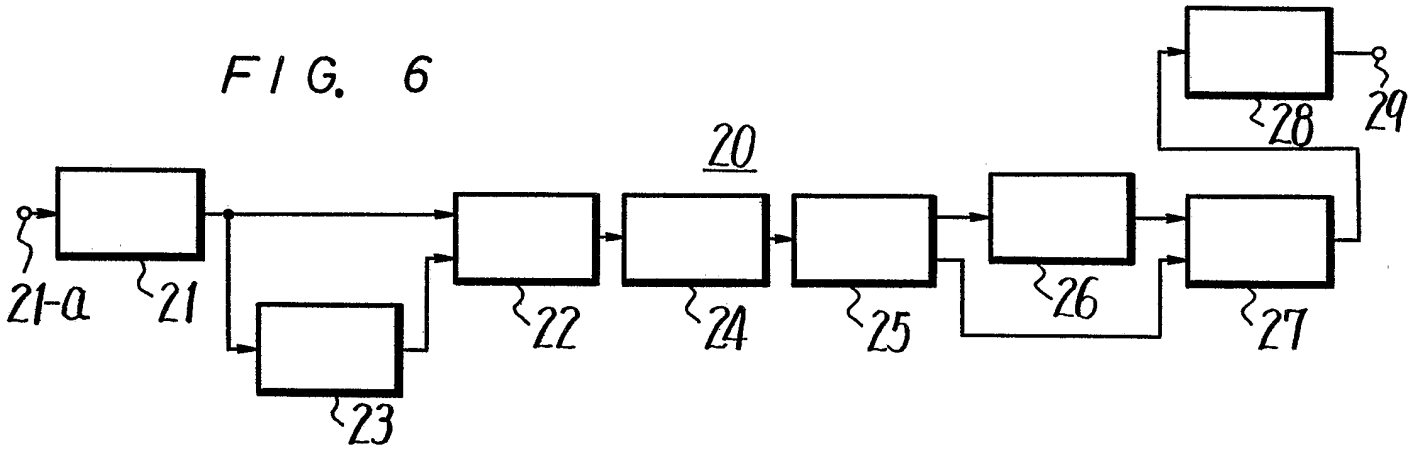



FIG. 6



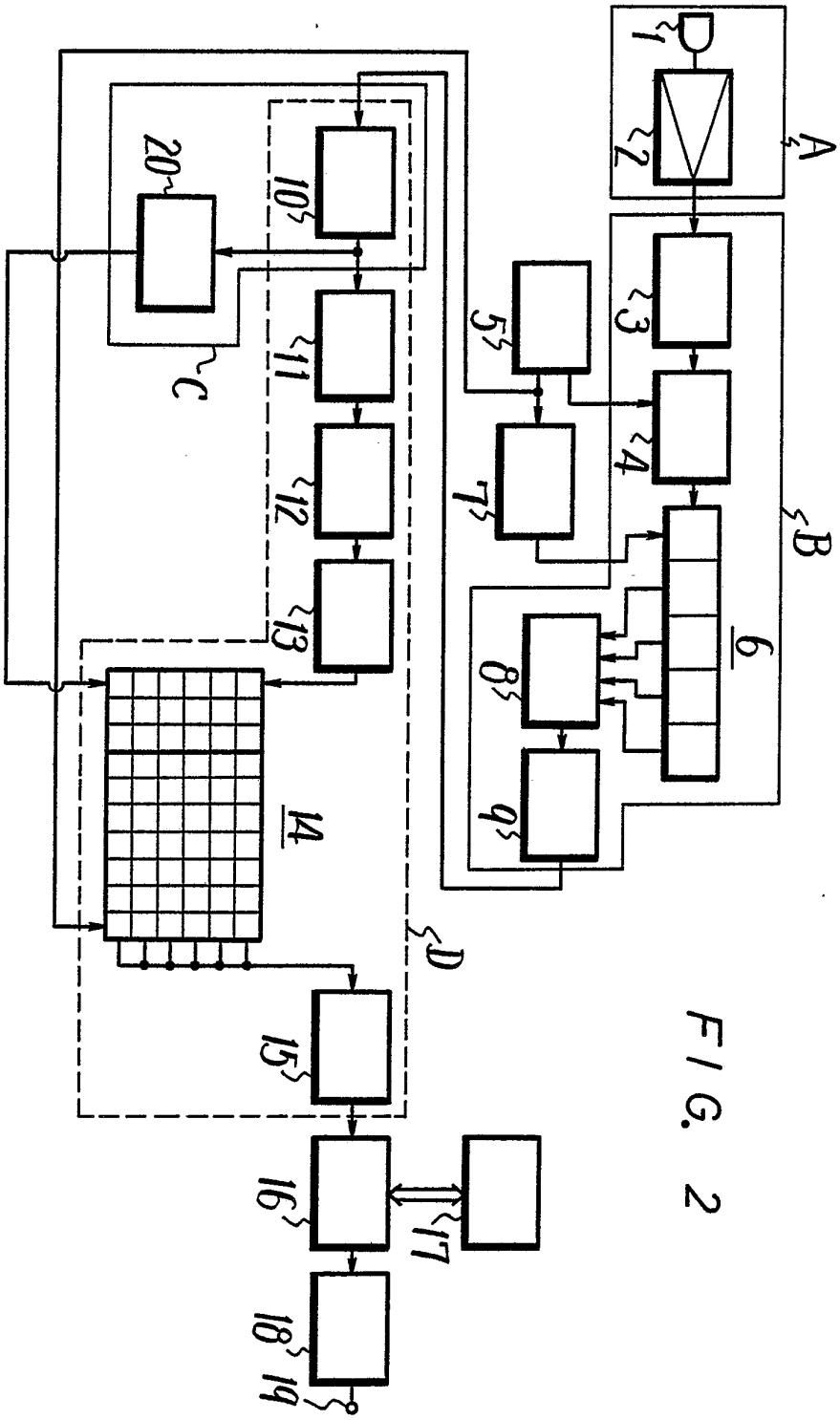



FIG. 2

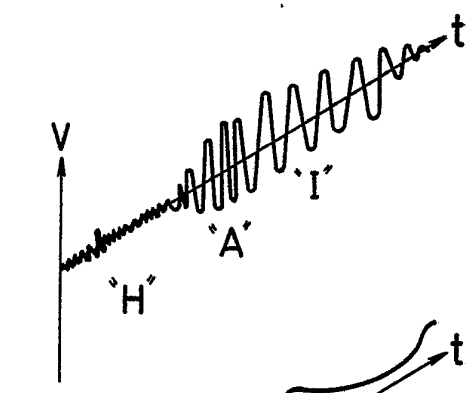



FIG. 3A

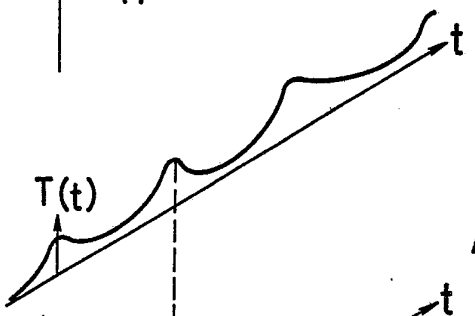



FIG. 3B

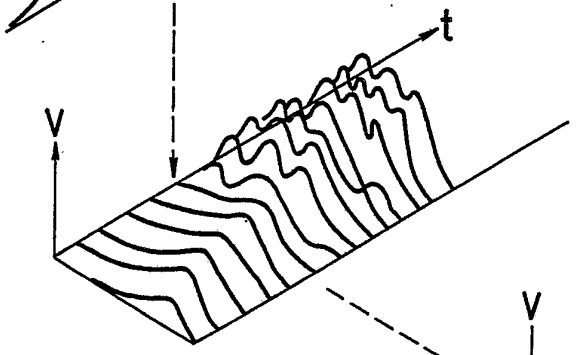



FIG. 3C


FIG. 3D

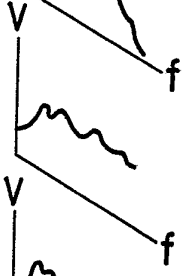



FIG. 3E

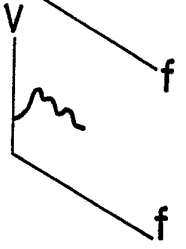



FIG. 3F

8300718

FIG. 3G

FIG. 3H

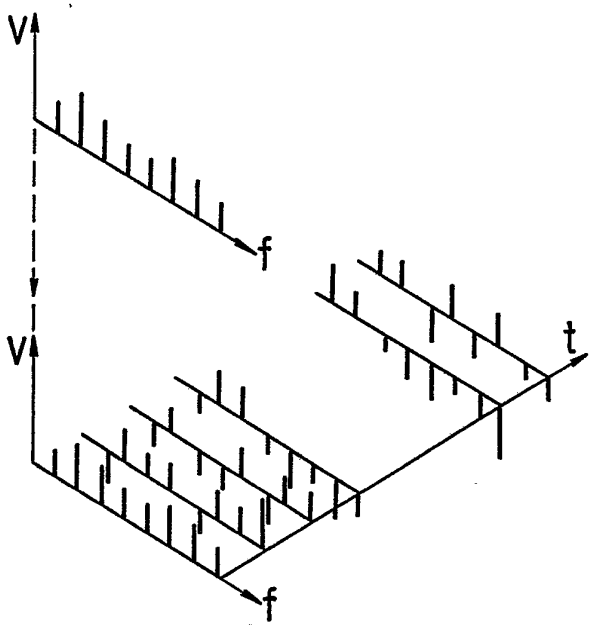



FIG. 4

| | | | | | | |
|---|---|---|---|---|---|---|
| HAI | ✳→H | H→A | A→I | I→✳ | | |
| IIE | ✳→I | I→E | E→✳ | | | |
| ZERO | ✳→Z | Z→E | E→R | R→O | O→✳ | |
| ICHI | ✳→I | I→✳ | ✳→CH | CH→I | I→✳ | |
| NI | ✳→N | N→I | I→✳ | | | |
| SAN | ✳→S | S→A | A→N | N→✳ | | |
| YON | ✳→Y | Y→O | O→N | N→✳ | | |
| GO | ✳→G | G→O | O→✳ | | | |
| ROKU | ✳→R | R→O | O→✳ | ✳→K | K→U | U→✳ |
| NANA | ✳→N | N→A | A→N | N→A | A→✳ | |
| HACHI | ✳→H | H→A | A→✳ | ✳→CH | CH→I | I→✳ |
| KYU | ✳→K | K→Y | Y→U | U→✳ | | |

FIG. 5A


FIG. 5B

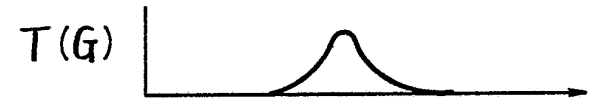



FIG. 5C


FIG. 5D


FIG. 5E


FIG. 5F

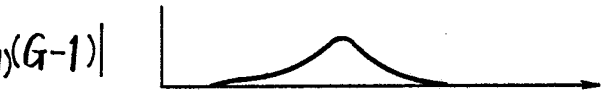



FIG. 5G

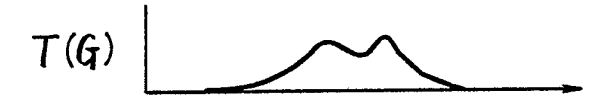



FIG. 5H

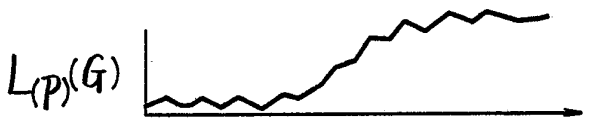



FIG. 5I

FIG. 7A

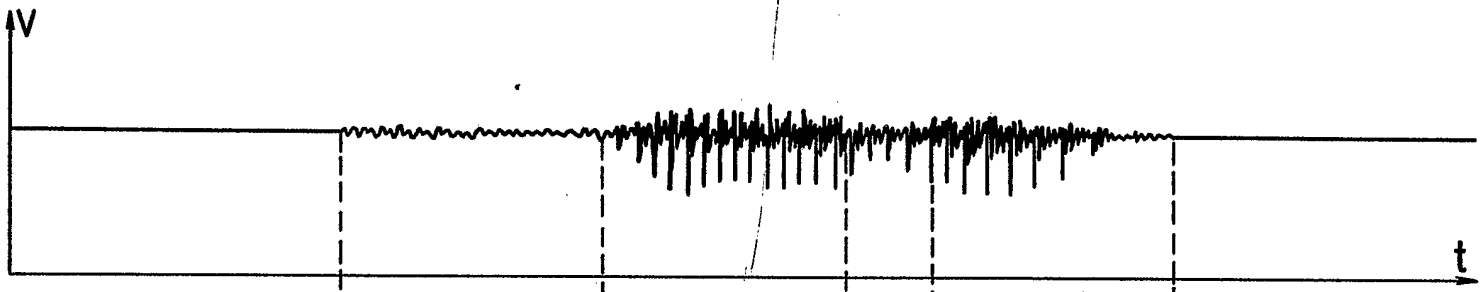



FIG. 7B

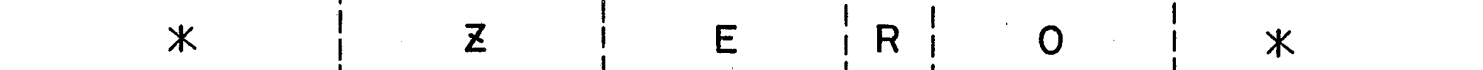



FIG. 7C