Octrooiraad

Nederland

(12) A Terinzagelegging (11) 8500377

(19) NL

(54) **Werkwijze en inrichting voor het segmenteren van spraak.**

(51) Int.Cl⁴.: G10L 5/00.

(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) Aanvrage Nr. 8500377.

Ingediend 12 februari 1985.

(32) - -

(33) - -

(31) - -

(62) - -

(43) Ter inzage gelegd 1 september 1986.

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

BAD ORIGINAL

N.V. Philips' Gloeilampenfabrieken te Eindhoven.

Werkwijze en inrichting voor het segmenteren van spraak.

De uitvinding heeft betrekking op een werkwijze voor het segmenteren van spraak, waarbij een akoestisch spraaksignaal wordt omgezet in N signalen, elk signaal behorende bij een tijdinterval i van N opvolgende tijdintervallen, waabij i looptvan 1 tot en met N. De uitvinding heeft eveneens betrekking op een inrichting voor het uitvoeren van de werkwijze.

Een dergelijke werkwijze respektievelijk inrichting wordt gebruikt voor het bepalen van (de grenzen van) segmenten waaruit de spraak is opgebouwd. Deze segmenten zijn klanken, bijvoorbeeld in de vorm van demi-syllaben, fonemen of difonen.

Het doel van een dergelijke handeling kan bijvoorbeeld zijn om op de verkregen resultaten een herkenningsprocedure toe te passen. Men spreekt in dit geval van woord- of spraakherkenning. Deze herkenningsprocedure kan inhouden dat men de verkregen segmenten vergelijkt met referentiepatronen van klanken. Bij voldoende overeenkomst tussen een segment en een referentiepatroon van een klank is het segment herkend.

Een andere mogelijkheid kan zijn dat men de verkregen segmenten gebruikt om een verzameling van allemaal verschillende klanken (bijvoorbeeld difonen) op te bouwen, waarna later met behulp van deze verzameling een synthese van kunstmatige spraak mogelijk is, zie bijvoorbeeld "Exploring the possibilities of speech synthesis with Dutch diphones" van B.A.G. Elsendoorn et al (1). Tot op heden wordt het opbouwen van een verzameling (of bibliotheek) van difonen door een getraind foneticus met de hand gerealiseerd en dit duurt dan ongeveer één jaar.

Een systeem voor het verkrijgen van difoonverzamelingen is bekend uit "Fabrication semi-automatique de dictionnaires de diphones" van M. Stella (2). Deze semi-automatische methode segmenteert slechts 72% van de difonen goed, zodat een operateur de resultaten naderhand interaktief dient te korrigeren.

Een segmentatiemethode die uitgaat van referentiepatronen van grotere klanten, te weten demi-syllaben, wordt beschreven in "A bootstrapping training technique for obtaining demisyllable

BAD ORIGINAL

reference patterns' van L.R. Rabiner et al (3).

Een nadeel van een dergelijke methode is dat, wil men nauwkeurige referentiepatronen afleiden, het opbouwen van een bibliotheek met dergelijke referentiepatronen erg veel tijd in beslag neemt. Vaak net zo veel tijd als op dit moment nodig is om op de bekende wijze een bibliotheek met difonen op te bouwen. Dit komt hoofdzakelijk doordat het aantal referentiepatronen voor een dergelijke bibliotheek erg groot is: voor de Nederlandse taal namelijk ongeveer 10.000 demisyllaben en 1800 difonen.

B.S. Atal (4) beschrijft in zijn publikatie "Efficient coding of LPC parameters by temporal decomposition" eveneens een segmentatiemethode. Deze methode heeft het nadeel dat het aantal te vinden segmenten niet vast ligt en niet bepaald wordt welke klank het afgesplitste segment is.

De uitvinding beoogt nu een werkwijze voor het segmenteren van spraak te verschaffen die nauwkeuriger en sneller werkt en waarbij bijgevolg in kortere tijd een verzameling van bijvoorbeeld difonen verkregen kan worden.

De werkwijze volgens de uitvinding heeft daartoe het kenmerk,

- dat voor opvolgende tijdintervallen i een funktie c_{ij} wordt afgeleid die een maat is voor de overeenkomst tussen het sigaan behorende bij het tijdinterval i en het signaal behorende bij het tijdinterval j, waarbij voor een zekere i j loopt van a_i ($a_i < i$) tot en met b_1 ($b_i > i$), en
- dat die waarden i, behorend bij een rangnummer i van een tijdinterval, worden bepaald die de begrenzingen vormen van gedeelten van het spraaksignaal waarvoor geldt dat de kurves c_{ij} behorend bij tijdintervallen liggend binnen een gedeelte ten minste ongeveer overeenkomen. Deze waarden i kunnen bijvoorbeeld worden bepaald doordat
- dat voor elk tijdinterval i het midden m_i van de bij het interval i behorende funktie c_{ij} wordt bepaald, en
- dat voor de genoemde waarden i die waarden i genomen worden die ten minste ongeveer overeenkomen met nuldoorgangen met eenzelfde teken in een funktie d_i, aangevende het verschil tussen het midden i en het rangnummer van het betreffende interval i. De waarden i komen overeen met de tippen gedurende een spraakuiting en geven de grenzen aan van de segmenten. Bepaalt men die waarden i die ten minste ongeveer overeenkomen met de nuldoorgangen van negatief naar positief in de funktie d_i

BAD ORIGINAL

dan verkrijgt men de foneemgrenzen. De segmenten zijn dan bijgevolg de fonemen. Een dergelijk segment wordt gekarakteriseerd doordat de parameters van de spraak, zoals formanten en bandbreedtes, binnen dit segment min of meer konstant zijn.

Een foneem kan gedefinieerd worden als zijnde de kleinste eenheid in een woord dat bij verandering een betekeniswijziging van het woord realiseert. Een difoon beschrijft de overgang tussen twee aan-

BAD ORIGINAL
8500377

grenzende fonemen in een woord. Later zal een afleiding voor difonen beschreven worden uitgaande van de op de hiervoor beschreven wijze verkregen fonemen.

Voor de funktie c_{ij} zou men de korrelatiefunktie kunnen nemen. Deze funktie heeft de waarde 1 voor gelijke tijdintervallen, dat wil zeggen i = j, en heeft een kleinere waarde naarmate de klanken op de tijdintervallen i en j meer verschillen.

a_i en b_i kunnen gelijk genomen worden aan die waarden i waarvoor geldt dat de funktie c_{ij} gelijk is aan een referentiewaarde c_r. De referentiewaarde kan daarbij gelijk aan p maal de maximale waarde van de funktie c_{ij} genomen worden, waarbij $p < 1$, bijvoorbeeld gelijk aan 0,8 is. Hierdoor wordt bereikt dat overeenkomstige segmenten in een spraakuitging elkaar niet beïnvloeden. Dat wil zeggen dat de bepaling van het midden m_i van het ene segment niet beïnvloed wordt (anders gezegd: verstoord wordt) door de aanwezigheid van een ander overeenkomstig segment verder of eerder in de spraakuiting.

Voor de N signalen kunnen de frekwentiespektra van het akoestische spraaksignaal in de N tijdintervallen genomen worden. Een andere mogelijkheid is dat voor de N signalen de spektrale omhullenden van het akoestische spraaksignaal in de N tijdintervallen wordt genomen.Het verkrijgen van spektrale omhullende bijvoorbeeld met behulp van de LPC-analyse, is op zich bekend uit "Linear prediction of speech" van J.D. Markel en A.H. Gray (5). Het gebruik van de spektrale omhullenden voor het afleiden van de funktie c_{ij} heeft het voordeel dat de invloed van de toonhoogte, wel aanweizg in de frekwentiespektra zelf, onderdrukt is.

De werkwijze zoals hiervoor beschreven levert dus een in segmenten onderverdeeldspraaksignaal op waarbij de segmentgrenzen worden bepaald door de waarden i, welke waarden weer de grenzen langs de tijdas voor het segment aangeven.

Het afleiden van een bibliotheek voor difonen zal men vaak realiseren door eerst uit een spraaksignaal segmenten in de vorm van fonemen te bepalen en uitgaande van het resultaat vervolgens de difonen af te leiden. Een op zich bekende spraakuiting, waarin dus een aantal van de af te leiden difonen zitten, wordt daartoe aan de hiervoor weergegeven segmentatieprocedure onderworpen. Men krijgt dan, indien men in de kurve $d_i = m_i - i$ de nuldoorgangen van negatief naar positief de segmentgrenzen die overeen zouden moeten komen met de foneem-

BAD ORIGINAL

grenzen van de fonemen waaruit de bekende spraakuiting is opgebouwd. Het kan zijn dat de segmentatieprocedure evenveel segmenten in de spraakuiting heeft bepaald als er fonemen zijn in de spraakuiting. In dat geval is precies aan te geven welk segment overeenkomt met welk foneem. Het kan echter ook zijn dat de segmentatieprocedure juist meer of minder segmenten oplevert dan het aantal fonemen waaruit de spraakuiting is opgebouwd. In deze gevallen is het niet direkt duidelijk welk segment overeenkomt met welk foneem.

In dat geval is een speciale toewijzingsprocedure nodig waarbij in het eerste geval (er zijn meer segmenten dan fonemen) er één of meer grenzen zullen moeten vervallen en in het tweede geval (er zijn minder segmenten dan fonemen) er één of meer grenzen zullen moeten worden toegevoegd. Om dit te kunnen realiseren heeft de werkwijze voor het segmenteren van het spraaksignaal van een op zich bekende spraakuiting in segmenten van klanken, bijvoorbeeld in de vorm van fonemen, het kenmerk, dat door middel van een op zich bekende tijdaanpassing tussen een aaneenschakeling van referentiepatronen van die klanken waaruit de spraakuiting is opgebouwd en het spraaksignaal van de spraakuiting een mate van overeenstemming wordt gerealiseerd die maximaal is, en dat die klanken aan die segmenten, begrensd door de waarden i, worden toegekend waarvoor de grenzen van klank en overeenkomstig segment de beste overeenstemming vertonen. Zijn er meer segmenten dan fonemen dan zijn er ergens in de spraakuiting twee naast elkaar liggende segmenten die tegenover één foneem liggen. De waarde i die de grens tussen deze twee segmenten bepaald wordt nu geschrapt. Zijn er meer fonemen dan segmenten dan zijn er ergens in de spraakuiting twee naast elkaar liggende fonemen die tegenover één segment liggen. In dat geval wordt het bewuste segment gesplitst door toevoeging van een waarde i overeenkomend met de grens tussen de twee fonemen.

Waren er evenveel segmenten als fonemen dan blijven de waarde i, zoals verkregen uit de funktie d_i, behouden.

Uitgaande van de nu verkregen waarden i voor de foneemgrenzen kunnen, gebruikmakend van omzettingsregels, de difoongrenzen worden bepaald. Dergelijk omzettingsregels zullen later worden besproken.

Een reden dat men de difoongrenzen (en dus de difonen) bepaald via eerst een afleiding van de foneemgrenzen is, dat men voor de voornoemde tijdaanpassing gebruik kan maken van de referentiepatronen van fonemen. Daar er slechts ongeveer 40 fonemen voor de Nederlandse taal

BAD ORIGINAL

zijn, hoeft men dus ook maar 40 referentiepatronen af te leiden en op te slaan, hetgeen goed en snel realiseerbaar is.

Speelt verder de vraag waarom men niet enkel uit de voornoemde tijdaanpassing van het spraaksignaal met de aaneenschakeling van referentiepatronen van de fonemen, de foneemgrenzen afleidt. Het antwoord hierop is dat een dergelijke methode erg gevoelig is voor onder andere coarticulatie en toevallige variaties in de spraak. Het gevolg is dat foneemgrenzen worden verkregen die een grote onnauwkeurigheid kunnen bezitten, zodat bijgevolg geen goede afleiding van de difoongrenzen respektieveljk difonen mogelijk is.

De voornoemde tijdaanpassing is dus alleen bedoeld om segmenten aan fonemen toe te wijzen en eventueel grenzen toe te voegen of weg te laten.

De onderverdeling van het spraaksignaal of spraakuiting in segmenten (bijvoorbeeld fonemen) kan ook gebruikt worden voor het herkennen van de spraakuiting. De werkwijze kan daartoe verder zijn gekenmerkt, doordat voor het herkennen van het spraaksignaal van een onbekende spraakuiting, elk segment wordt vergeleken met referentiepatronen van op zich bekende klanken, en dat die klank aan een segment wordt toegekend waarvoor de mate van overeenstemming tussen dit segment en de klank maximaal is. In het bijzonder indien de spraakuiting is gesegmenteerd in segmenten in de vorm van fonemen.

In dat geval hoeft men eveneens slechts de referentiepatronen van de veertig fonemen op te slaan en de segmenten in de onbekende spraakuiting elk dus te vergelijken met maximaal veertig referentiepatronen. Is de maximale overeenstemming bepaald dan geeft de aaneenschakeling van de referentiepatronen aan uit welke fonemen de onbekende spraakuiting is opgebouwd.

Een inrichting voor het uitvoeren van de werkwijze, met een ingangsklem voor het ontvangen van een elektrische signaal overeenkomend met het akoestische spraaksignaal, een eerste eenheid met een ingang gekoppeld met de ingangsklem en een uitgang, welke eerste eenheid is ingericht voor het afleiden van de N signalen en voor het toevoeren van deze N signalen aan zijn uitgang, heeft het kenmerk, dat de inrichting verder is voorzien van:

een tweede eenheid met een ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke tweede eenheid is ingericht voor het afleiden van de funktie c_{ij} voor opvolgende intervallen i

BAD ORIGINAL

uit de signalen behorende bij de tijdintervallen i en j en voor het toevoeren van de funktie c_{ij} aan zijn uitgang, en

- een deelinrichting met een ingang gekoppeld met de uitgang van de tweede eenheid en een uitgang, welke deelinrichting is ingericht voor het bepalen van die waarden i die de begrenzingen vormen van gedeelten van het spraaksignaal waarvoor geldt dat de kurves c_{ij} behorend bij tijdintervallen liggend binnen een gedeelte ten minste ongeveer overeenkomen. De deelinrichting kan daarbij zijn voorzien van :
- een derde eenheid met een ingang gekoppeld met de ingang van de deelinrichting en een uitgang, welke derde eenheid is ingericht voor het voor elk interval i afleiden van het midden m_i van de bijbehorende funktie c_{ij} en voor het toevoeren van de waarden m_i aan zijn uitgang, en
- een vierde eenheid met een ingang gekoppeld met de uitgang van de derde eenheid en een uitgang gekoppeld met de uitgang van de deelinrichting, welke vierde eenheid is ingericht voor het bepalen van die waarden i die ten minste ongeveer overeenkomen met nuldoorgangen in de funktie d_i en voor het toevoeren van die waarden i aan zijn uitgang.

Met deze inrichting kan een segmentatie van het spraaksignaal in segmenten waarbinnen de parameters van de spraak min of meer konstant zijn, gerealiseerd worden. De waarden i geven de onderlinge begrenzingen van de segmenten aan in de vorm van een rangnummer van een tijdinterval. Bijgevolg komen deze waarden i overeen met tijdstippen gedurende het spraaksignaal (de spraakuiting)

De vierde eenheid kan verder zijn ingericht voor het bepalen van die waarden i die ten minste ongeveer overeenkomen met de nuldoorgangen van negatief naar positief in de funktie d_i, waarbij $d_i = m_i - i$. De segmenten zijn in dit geval in de vorm van fonemen.

Een inrichting voor het uitvoeren van de werkwijze voor het segmenteren van een op zich bekende spraakuiting in segmenten van klanken, bijvoorbeeld in de vorm van fonemen, kan zijn gekenmerkt, door de inrichting verder is voorzien van een geheugen voor het opslaan van de referentiepatronen van klanken, een vijfde eenheid met een eerste ingang gekoppeld met een uitgang van het geheugen, een tweede ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke vijfde eenheid is ingericht voor het door middel van een op zich bekende tijdaanpassing tussen een aaneenschakeling van referentiepatronen van die klanken waaruit de op zich bekende spraakuiting is opgebouwd en het

BAD ORIGINAL

spraaksignaal van de spraakuiting, realiseren van een maximale mate van overeenstemming en voor het leveren van de op deze wijze verkregen grenzen aan zijn uitgang, en een kombinatie-eenheid met een eerste ingang gekoppeld met de uitgang van de vierde eenheid en een tweede ingang gekoppeld met de uitgang van de vijfde eenheid en een uitgang, welke vijfde eenheid is ingericht voor het toekennen van die klanken aan die segmenten begrensd door de waarden i, waarvoor de grenzen van klank en overeenkomstig segment de beste overeenstemming vertonen. De kombinatie-eenheid kan daarbij verder zijn ingericht voor het toevoegen van en het weglaten van één of meer waarden i en wel net zoveel waarden i als er minder respektievelijk meer segmenten zijn dan overeenkomstige klanken en voor het leveren van de al dan niet veranderde groep van waarden i aan zijn uitgang. Hiermee wordt bereikt dat, uitgaande van referentiepatronen van slechts 40 à 50 fonemen, aan de in de vierde eenheid verkregen segmenten elk een zeker foneem kan worden toegekend, en waarbij bovendien gekorrigeerd kan worden indien het aantal segmenten kleiner is dan of groter is dan het aantal fonemen waaruit de bekende spraakuiting is opgebouwd.

De inrichting kan voor het verkrijgen van de difonen verder zijn voorzien van een zesde eenheid, met een ingang gekoppeld met de uitgang van de kombinatie-eenheid en een uitgang, welke zesde eenheid is ingericht voor het, uitgaande van de waarden i van de kombinatie-eenheid voor de klankgrenzen en gebruikmakend van omzettingsregels, afleiden van de difoongrenzen voor het toevoeren van deze difoongrenzen aan zijn uitgang.

Voor het mogelijk maken van het herkennen van het spraaksignaal van een onbekende spraakuiting kan de inrichting zijn gekenmerkt, doordat de inrichting verder is voorzien van een geheugen voor het opslaan van referentiepatronen van op zich bekende klanken en van een vijfde eenheid met een eerste ingang gekoppeld met de uitgang van het geheugen, een tweede ingang gekoppeld met de uitgang van de vierde eenheid en een derde ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke vijfde eenheid is ingericht voor het vergelijken van elk segment met referentiepatronen van op zich bekende klanken, en voor het toekennen van die klank aan een segment waarvoor de mate van overeenstemming tussen dit segment en de klank maximaal is, en is ingericht voor het leveren aan zijn uitgang van informatie betreffende de aaneenschakeling van referentiepatronen van die klanken die elk maximaal over-

BAD ORIGINAL

eenstemmen met het segment waaraan zij zijn toegekend.

De uitvinding aan de hand van de hierna volgende figuurbeschrijving zal nader worden uiteengezet. Hierin toont:

Fig. 1 een eerste uitvoeringsvoorbeeld van de inrichting volgens de uitvinding,

Fig. 2 een tweede uitvoeringsvoorbeeld van de inrichting volgens de uitvinding,

Fig. 3 in fig. 3a het elektrische signaal van de spraakuiting als funktie van de tijd, in fig. 3b de kurve d_i verkregen in de segmentatieprocedure en in figuuur 3c de fonemen toegekend aan de verkregen segmenten,

Fig. 4 de N spektrale omhullenden van het spraaksignaal van fig. 3a voor opvolgende tijdintervallen $i(1 \leqslant i \leqslant N)$,

Fig. 5 het frekwentiesprektrum van een spraaksignaal in een zeker tijdinterval en het bijbehorende omhullende frekwentiespektrum,

Fig. 6 de funktie c_{ij} voor een zeker tijdinterval $i(1 \leqslant j \leqslant N)$,

Fig. 7 de funkties c_{ij} voor opvolgende tijdintervallen i,

Fig. 8 in fig. 8a de funktie $c_{i,px}$, in fig. 8b het oorspronkelijke tijdsignaal van de spraakuiting, en in fig. 8c weer de funktie d_i met daarin de difoongrenzen 9 en de segmentgrenzen i,

Fig. 9 in fig. 9a en 9b het toevoegen van een additionele segmentgrens en in fig. 9c en 9d het weglaten van een segmentgrens, en

Fig. 10 de tijdaanpassing tussen de spraakuiting en de referentiepatronen van de fonemen.

Verder bevat de figuurbeschrijving drie tabellen met omzettingsregels van foneemgrenzen naar difoongrenzen voor de Nederlandse, Duitse en Engelse taal.

Figuur 1 toont een eerste uitvoeringsvoorbeeld van de inrichting volgens de uitvinding. De inrichting wordt gebruikt voor het afleiden van een difoonverzameling. Daartoe wordt door een geoefende spreker een op zich bekende spraakuiting via een mikrofoon 1 aan de inrichting aangeboden. De spreker dient er daarbij op te letten dat hij met een zekere snelheid spreekt. Deze snelheid mag niet te groot of te laag zijn. Bovendien moet bij voorkeur met een vaste toonhoogte spreken. De spraakuitingen die worden gebruikt voor het verkrijgen van een difoonverzameling zijn meestal betekenisloze woorden, zoals bijvoorbeeld het woord "nenoone". Dit woordt is opgebouwed uit de fonemen "n", "e", "n", "oo", "n" en "e". Hieruit kunnen de volgende difonen worden

BAD ORIGINAL

gedestilleerd, te weten "ne", "en", "noo", "oon" en "ne". Doordat difonen in feite de overgangen van twee fonemen aangeven is het duidelijk dat er veel meer difonen zijn dan fonemen, voor de Nederlandse taal ongeveer 1800 difonen tegen ongeveer 40 à 50 fonemen. De mikrofoon 1 is via de ingangsklem 2 gekoppeld met de analoog-digitaal omzetter 3. Hierin wordt het elektrische signaal bemonsterd met een bemonsterfrekwentie van ongeveer 10 kHz en gedigitaliseerd. Daar de informatie voor spraak zich voornamelijk in een frekwentiegebied beneden 5 kHz bevindt, is de bemonsterfrekwentie van 10 kHz voldoende. Fig. 3a toont het gedigitaliseerde tijdsignaal van het woord "nenoone" dat verschijnt aan de uitgang van analoog-digitaal-omzetter 3. Een eerste eenheid 4 is met een ingang 5 gekoppeld met de uitgang van de A/D omzetter 3. Deze eenheid 4 is ingericht voor het afleiden van N signalen $S_i(f)$ $(1 \leqslant i \leqslant N)$. Deze N signalen worden verkregen door het tijdsignaal van fig. 3a in N intervallen te verdelen en voor elk interval bijvoorbeeld het frekwentiespektrum van de spraakuiting in het betreffende interval te nemen.

Een andere mogelijkheid is dan men de spektrale omhullende van de spraakuiting in het betreffende interval neemt. Hiervoor zou men de LPC-analyse kunnen gebruiken. Stel dat men intervallen van 10 ms lang neemt dan wordt de bijbehorende spektrale omhullende berekend uit een tijdsignaal bestaande uit 100 bemonsteringen. Fig. 4 toont de eerste vijftig spektrale omhullenden $S_i(f)$ voor de eerste 500 ms. van het spraaksignaal.

Fig. 5 toont met de kurve $S_i'(f)$ het frekwentiespektrum van de spraakuiting voor het interval i. Dit spektrum kan men door middel van een snelle fouriertransformatie verkrijgen. Wat opvalt in de kurve $S_i'(f)$ zijn de pieken op $f = f_o$ en veelvouden daarvan. De frekwentie f_o geeft de toonhoogte aan van het spraaksignaal. De kurve $S_i(f)$ geeft de met behulp van LPC-analyse verkregen spektrale omhullende aan die over een aantal dB naar boven verschoven is. De pieken in deze kurve bij de frekwentiewaarden f_1, f_2, f_3 en f_4 geven de eerste vier formanten aan. Met de LPC-technieken worden in feite alle parameters bepaald van een filter dat is opgebouwd uit een serieschakeling van een aantal (in dit geval vier) tweede orde filters. De betreffende parameters zijn dan de resonatiefrekwenties van deze filters (welke overeenkomen met de voornoemde formantfrekwenties) en de bandbreedtes, zie de dissertatie van Vogten (6), in het bijzonder § 2.1.

Met deze parameters is het vervolgens achteraf mogelijk om

BAD ORIGINAL

spraak te synthetiseren, zie de publikatie "MEA 8000 voice synthesizer, principles and interfacing " (7), in het bijzonder het gedeelte "Principles of formant speech synthesis" de pagina's 2-4.

Een tweede eenheid 7 is met zijn ingang 8 gekoppeld met de uitgang 6 van de eerste eenheid 4. Gebruikmakend van de N signalen $S_i(f)$, zie fig. 4, die door de eerste eenheid 4 aan haar uitgang 6 worden aangeboden, worden in de tweede eenheid 7 voor elk interval i een funktie c_{ij} afgeleid. j loopt hierbij van a_i (waarbij $a_i < i$) naar b_i (waarbij $b_i > i$). Fig. 6 toont een dergelijke kurve c_{ij} en fig. 7 toont voor een aantal opvolgende tijdintervallen i de kurves c_{ij}, waarbij c_{ij} is berekend volgens de volgende formules:

$$R_{ij} = \int_0^{5\ kHz} S_i(f)S_j(f)df \qquad (1)$$

en

$$c_{ij} = \frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}} \qquad (2)$$

Uit formule (2) blijkt dat c_{ij} de waarde 1 bezit voor j = i, zie fig. 6. De kurve c_{ij} geeft aan de mate van overeenstemming van de spektrale omhullende voor het tijdinterval i en die voor het tijdinterval j. Uit fig. 7 is duidelijk dat de spektrale omhullenden voor de tijdintervallen 10 tot en met 17 veel overeenstemming met elkaar hebben. Hetzelfde geldt voor de tijdintervallen 19 tot en met 27. Ook in fig. 4 is dit redelijk goed te zien. Er zijn delen in het spraaksignaal waarin de spektrale omhullenden veel op elkaar lijken, gevolgd door een overgang naar een volgend stationair deel. Elk deel geeft dus aan dat gedurende zo'n deel de bovengenoemde parameters, zoals de formanten en hun bijbehorende bandbreedtes min of meer konstant zijn in de tijd.

Een derde eenheid 10 is met zijn ingang gekoppeld met de uitgang 9 van de tweede eenheid 7. De derde eenheid 10 krijgt van de tweede eenheid 7 de kurves c_{ij} toegevoerd en bepaalt voor de kurve c_{ij} behorend bij elk interval het midden m_i van deze kurve, bijvoorbeeld op de volgende wijze:

BAD ORIGINAL

8500377

$$m_i = \frac{\int_{a_i}^{b_i} jC_{ij}\,dj}{\int_{a_i}^{b_i} C_{ij}\,dj} \qquad (3)$$

De waarden a_i en b_i kunnen redelijk willekeurig gekozen worden. Bij voorkeur zal men die waarden j voor a_i en b_i nemen waarvoor geldt dat $C_{i,a_i} = C_{i,b_i} = C_r$, zie fig. 6, waarbij C_r gelijk is aan $p.c_{ii}$ (c_{ii} zijnde de maximale waarde in de kurve c_{ij}, in dit geval gelijk aan 1. Bij voorkeur neemt men p gelijk aan 0,8 of 0,9. Door deze waarden voor a_i en b_i te kiezen krijgt men een nauwkeuriger bepaling van de waarde m_i dan indien men de integratie in formule (3) zou laten uitstrekken van 1 tot N, dus over de gehele spraakuiting. In dit laatste geval zouden min of meer stationaire gedeelten elders in de spraakuiting die een grote gelijkenis hebben met het gedeelte waarvoor het midden bepaald dient te worden, de bepaling van dat midden ernstig verstoren. Als voorbeeld: in het woord "nenoone" komen driemaal min of meer konstante gedeelten voor die overeenkomen met het foneem "n". Wil men het midden m_i bepalen in het konstante gedeelte van de middenste "n" dan dient men het integratiegebied niet zo groot te kiezen dan een van de andere fonemen "n" er binnen valt.

In fig. 7 is met een vertikale lijn de waarde j aangegeven overeenkomende met het midden m_i voor elk interval. Duidelijk zichtbaar is dat het midden binnen een stationair gedeelte, $10 < j < 17$ en $20 < j < 27$, praktisch niet verandert voor opvolgende i.

Een vierde eenheid 13 is met zijn ingang 14 gekoppeld met de uitgang 12 van de derde eenheid 10. De vierde eenheid 13 bepaalt in het blok 15 de grootheid d_i uit de gegevens verkregen in de derde eenheid 10.

Voor de funktie d_i geldt:

$$d_i = m_i - i \qquad (4)$$

BAD ORIGINAL

In figuur 7 is d_i voor ieder interval i aangegeven door middel [illegible] Verder toont fig. 3b

de kurve d_i als funktie van i. De nuldoorgangen van positief naar negatief in de funktie d_i komen ten minste ongeveer overeen met de middens m_i in het bijbehorende stationaire deel. De nuldoorgangen van negatief naar positief, aangeduid met i_1 tot en met i_m, geven de posities van de overgangen tussen de stationaire gedeelten (fonemen) aan. Deze nuldoorgangen worden bepaald in het blok 16 van de vierde eenheid 13.

Uit fig. 3b in kombinatie met fig. 3c wordt duidelijk dat het eerste en het laatste segment SI overeenkomen met de stilte net voor en direkt na de spraakuiting. Vervolgens blijven er nog zes segmenten over die precies overeenstemmen met de zes fonemen waaruit het woord "nenoone" is opgebouwd, Aan de zes segmenten is dus direkt het juiste foneem toe te wijzen.

Soms kan het voorkomen, dat met behulp van de hiervoor beschreven segmentatieprocedure, er meer of juist minder segmenten verkregen worden dan het aantal fonemen waaruit de spraakuiting is opgebouwd. In dat geval kan aan een segment niet direkt een zeker foneem worden toegekend. Om ook in die gevallen op een korrekte wijze fonemen aan segmenten toe te kunnen kennen is de inrichting verder voorzien van een geheugen 18. In dit geheugen zitten opgeslagen voor elk foneem x: één referentiepatroon van het foneem en de minimale en maximale tijdsduur ($t_{x,\,min}$ en $t_{x,max}$) van het foneem. Dit referentiepatroon bestaat uit de formantwaarden en de bijbehorende bandbreedtes. Uitgaande van deze waarden is volgens de LPC analyse een spektrale omhullende S(f) te berekenen, zoals in fig. 5 aangegeven. Deze spektrale omhullende is de omhullende behorend bij een zeker foneem x, en zal in het vervolg worden aangeduid met $S_{px}(f)$.

Analoog aan de formules (1) en (2) kan nu de grootheid $c_{i,px}$ worden afgeleid aangevende de overeenkomst tussen de spektrale omhullende van het tijdsignaal in interval i en van het foneem x. Dit gebeurt in de vijfde eenheid 19. Deze eenheid heeft daartoe een eerste ingang 20, gekoppeld met het geheugen 18 voor het toevoeren aan de eenheid 19 van de referentiepatronen waaruit de omhullenden $S_{px}(f)$ kunnen worden afgeleid, en een tweede ingang 21 gekoppeld met de uitgang 6 van de eerste eenheid 4, voor het toevoeren van de omhullenden $S_i(f)$ behorende bij de tijdintervallen i. In de vijfde eenheid 19 worden de grootheden $c_{i,px}$ afgeleid en wordt er een tijdaanpassing gepleegd tussen de spraakuiting en een aangenschakeling van referentiepatronen van fonemen, dat de mate van overeenstemming tussen beide signalen maximaal is. Fig. 8a toont

BAD ORIGINAL

het resultaat van deze tijdaanpassing voor de spraakuiting "nenoone", weergegeven in fig. 8b. Voor x = 1 in fig. 8a geldt de stilte voor de spraakuiting, x = 2 komt overeen met het foneem "n", x = 3 met het foneem "e", x = 4 weer met het foneem "n", x = 5 met het foneem "oo", x = 6 met het foneem "n", x = 7 weer met het foneem "e" en x = 8 weer met de stilte.

Doordat het hier een bekende spraakuiting betreft, weet men reeds welke fonemen aanwezig zijn en in welke volgorde ze moeten liggen. Het enige waar het om gaat is het bepalen van de grenzen i_1' tot en met i_7'.

Deze grenzen kunnen als volgt bepaald worden. In fig. 10 staat langs de horizontale as de tijd uit, aangeduid in het rangnummer i van de intervallen i. Er geldt natuurlijk dat $1 \leqslant i \leqslant N$. Langs de vertikale as staan de (referentiepatronen van de) fonemen p_1 tot en met p_X uit. X is gelijk 8 voor het geval de stilten voor en na de spraakuiting ook als fonemen meegenomen worden. De foneem p_x bevindt zich bij de tijdintervallen $i_{x-1}' + 1$ tot en met i_x'. Op de lijn l die van linksonder naar rechtsboven in de grafiek loopt zijn steeds de waarden $c_{i,\,px}$ uitgezet.

Men bepaalt nu de waarden i_x' (waarbij $1 \leqslant x < X$) zodanig dat de funktie:

$$\sum_{x=1}^{X} \sum_{i'_{x-1}+1}^{i_x'} c_{i,px} \qquad \text{waarbij } i_o' = 0 \text{ en } i_X' = N,$$

maximaal is. Deze maximalisatie kan uitgevoerd worden met behulp van dynamische programmering, waarbij voorwaarden gesteld kunnen worden aan i_x', zoals:

$$t_{x,min} \leqslant i_x' - i_x' - 1 \leqslant t_{x,max}$$

Het dynamisch programmeren is een op zich bekende techniek: zie "Dynamic programming" van R. Bellman (8) en de publikatie van Ney (9) "Dynamic programming as a technique for pattern recognition". Op de hiervoor weergegeven wijze worden alle grenzen i_1' tot en met i_7' bepaald. Vervolgens worden de grenzen i_1 tot en met i_7 verkregen uit de segmentatieprocedure, en de grenzen i_1' tot en met i_7' verkregen

BAD ORIGINAL

[illegible] kombinatie-eenheid 23. Daartoe

is een eerste en een tweede ingang 24 respektievelijk 25 van de kombinatie-eenheid 23 gekoppeld met een uitgang 17 van de vierde eenheid 13 respektievelijk met een uitgang 22 van de vijfde eenheid 19.

Deze kombinatie-eenheid 23 is ingericht voor het toekennen van die fonemen verkregen volgens de tijdaanpassing in de eenheid 19 (zie fig. 8a), aan die segmenten begrensd door de waarden i, verkregen volgens de segmenteringsmethode in de eenheid 13 (zie fig. 3b en ook fig. 8c die dezelfde kurve vertoont), waarvoor de grenzen van foneem en overeenkomstig segment de beste overeenstemming vertonen.

Uit fig. 8a en c is duidelijk dat de grenzen i_1 en i_1' bij elkaar horen. Hetzelfde geldt voor de grenzen i_2 en i_2', i_3 en i_3', i_4 en i_4', i_5 en i_5', i_6 en i_6' en i_7 en i_7'. De kombinatie-eenheid 23 levert dan ook in dit geval aan zijn uitgang de waarden i_1 tot en met i_7.

Het had ook zo kunnen zijn dat de segmentatieprocedure bijvoorbeeld één segment minder had opgeleverd dan het aantal fonemen waaruit de spraakuiting is opgebouwd. Dit toont fig. 9a en 9b. Fig. 9a toont de resultaten van de segmentatieprocedure. Uit de kurve d_i, die hier niet verder is aangegeven, zijn de grenzen i_1 tot en met i_4 verkregen terwijl de spraakuiting uit zes fonemen is opgebouwd. Uit de methode van de tijdaanpassing volgen de grenzen i_1' tot en met i_5', zoals in fig. 9b is aangegeven. De kurve $C_{i,\,px}$ is hier niet aangegeven. Duidelijk is dat er overeenstemming bestaat tussen de grenzen i_1 en i_1', i_2 en i_2', i_3 en i_4' en i_4 en i_5'. Er zal in dit geval door de kombinatie-eenheid 23 een extra grens bijgevoegd worden overeenkomende met i_3'.

Een andere mogelijkheid is dat de segmentatieprocedure juist één segment meer had opgeleverd dan het aantal fonemen waaruit de spraakuiting is opgebouwd, Dit toont fig. 9c en 9d. Fig. 9c toont de resultaten van de segmentatieprocedure: vijf grenzen i_1 tot en met i_5. Fig. 9d toont de resultaten van de tijdaanpassing: vier grenzen i_1' tot en met i_4'. Duidelijk is dat er overeenstemming bestaat tussen de grenzen i_1 en i_1', i_2 en i_2', i_4 en i_3' en i_5 en i_4'. Door de kombinatie-eenheid 23 zal in dit geval de grens i_3 weggelaten worden.

De aldus verkregen en eventueel gekorrigeerde grenzen worden vervolgens toegevoerd aan de zesde eenheid 27. Deze eenheid 27 bepaalt, uitgaande van de door de kombinatie-eenheid 23 verkregen grenzen i en gebruikmakend van omzettingsregels zoals aangegeven in tabel I de difoongrenzen g_1 tot en met g_8, zie fig. 8b en 8c. De tabel I toont enkel de omzettingsregels voor de belangrijkste en meest voorkomende difonen. De

BAD ORIGINAL

difoon liggend tussen g_1 en g_2 beschrijft de overgang van de stilte naar de "n", de difoon liggend tussen g_2 en g_3 de overgang "ne", die liggend tussen g_3 en g_4: "en", die tussen g_4 en g_5: "noo", die tussen g_5 en g_6: "oon", die tussen g_6 en g_7: "ne" en die tussen g_7 en g_8 de overgang tussen de "e" en de stilte. Uit de tabel I nr. 31 is duidelijk dat g_2, g_4 en g_6 in het midden van de segmenten bepaald door i_1, i_2; i_3, i_4 en i_5, i_6 (moeten) liggen. Uit de tabel I nr. 12 is duidelijk dat g_3 en g_7 6 tijdintervallen na i_2 respektievelijk i_6 genomen moet worden. Uit de tabel I nr. 8 volgt dat g_5 6 intervallen na i_4 langs de i-as ligt en uit de tabel I nr. 17 volgt dan g_1 en g_8 in het midden van het eerste en het laatste segment moeten liggen.

De fonemen uit de tabel I onder nr. 19 tot en met 24 zijn de zogenaamde plofklanken en bestaan uit twee gedeelten. Een eerste gedeelte komt overeen met de stilte vóór de plof. Het tweede gedeelte is de plof zelf in de tabel "burst" genoemd. De hiervoor beschreven segmentatie-procedure zal deze fonemen in het algemeen in twee segmenten, overeenkomende met deze twee gedeelten, segmenteren.

De bijbehorende difoongrens wordt nu volgens de tabel I gekozen op 1 tijdinterval liggend vóór het begin van het segment overeenkomende met de plof.

De "glottal stop" komt overeen met de stilte die wordt toegevoegd tijdens het uitspreken van losse letters, bijvoorbeeld bij het uitspreken van de naam van de spraaksynthetisator "MEA 8000", eerder vermeld.

Verder vindt men in de tabellen II en III omzettingsregels voor het afleiden van de belangrijkste Duitse en Engelse difonen.

De op de hiervoor beschreven wijze verkregen difonen worden vervolgens opgeslagen in een geheugen 30 voor het vormen van een bibliotheek van difonen. Het geheugen 30 is daartoe met twee ingangen gekoppeld met de uitgangen 6 en 29 van de eerste respektievelijk zesde eenheid 4 respektievelijk 27. In het geheugen 30 worden voor elk difoon en voor elk tijdinterval liggend binnen de grenzen g van het difoon opgeslagen

- de eerdergenoemde vier (of vijf) formanten en de bijbehorende bandbreedtes,
- een amplitudeparameter die samenhangt met de luidheid (ook afgeleid in de eerste eenheid 4),
- een stemhebbend-stemloosbeslissing (te weten of het een periodiek signaal of een ruisvormig signaal is, eveneens

BAD ORIGINAL

afgeleid in de eerste eenheid 4).

Met deze gegevens is het realiseren van synthetische spraak achteraf mogelijk, zie de dissertatie van Vogten (6).

Fig. 2 toont een tweede uitvoeringsvoorbeeld van de inrichting volgens de uitvinding. De inrichting wordt gebruikt voor het herkennen van een onbekende spraakuiting. Gelijke referentiecijfers in de figuren 1 en 2 stellen dezelfde elementen voor. De onbekende spraakuiting wordt aangeboden aan de mikrofoon 1 en in de A/D omzetter 3 omgezet in een digitaal signaal. De eerste eenheid 4 levert weer de funkties $S_i(f)$, zie fig. 4. De tweede eenheid 7 bepaalt de funkties c_{ij}, zie fig. 7. De derde eenheid 10 bepaalt vervolgens de middens m_i, zie eveneens fig. 7. De vierde eenheid 13 bepaalt de kurve d_i en de waarden i_1, ..., i_M, zie fig. 3.

We hebben nu weer een aantal segmenten gekregen die overeenstemmen met in principe evenzoveel referentiepatronen van de bij die segmenten behorende fonemen waaruit de onbekende spraakuiting is opgebouwd, Het geheugen 18 bevat daartoe weer de referentiepatronen van alle fonemen. De vijfde eenheid 31 is hier ingericht voor het vergelijken van elk segment met de referentiepatronen van de fonemen en voor het toekennen van dat foneem aan een segment waarvoor geldt dat de mate van overeenstemming tussen dit segment en het foneem maximaal is. De vijfde eenheid heeft daartoe een eerste, een tweede en een derde ingang 32, 33 respektievelijk 35 gekoppeld met het geheugen respektievelijk met de uitgang 17 van de vierde eenheid, respektievelijk de uitgang 6 van de eerste eenheid 4, en een uitgang 34. Aan deze uitgang 34 verschijnt de informatie betreffende de aaneenschakeling van referentiepatronen van die fonemen die elk maximaal overeenstemmen met het segment waaraan zij zijn toegekend. De onbekende spraakuiting is daarmee herkend.

De vergelijking van een segment met een referentiepatroon vindt bijvoorbeeld als volgt plaats.

Voor ieder segment en een referentiepatroon van een zeker foneem x wordt de kurve $c_{i,\,px}$ berekend. Voor een segment uit de spraakuiting en het referentiepatroon van dat foneem x dat ook werkelijk in de spraakuiting is uitgesproken verkrijgt men een kurve $c_{i,\,px}$ zoals in fig. 8a door de kurve tussen bijvoorbeeld i_3' en i_4' is aangegeven. Er zij hierbij wel vermeld dat de grenzen niet bij i_3' en i_4' kunnen liggen doch bij i_3 en i_4 verkregen uit de kurve d_i. De vergelijking vindt dus plaats tussen enerzijds het referentiepatroon van een foneem en een

BAD ORIGINAL

segment begrensd door de waarden i verkregen uit de vierde eenheid 13, dat wil zeggen i_1 tot en met i_M.

Bij vergelijking tussen een segment van de spraakuiting en een ander referentiepatroon zal een $c_{i,\,px}$ verkregen worden met een veel lagere waarde. Vervolgens wordt aan een segment dat foneem x toegewezen waarvoor geldt dat $c_{i,\,px}$ (geïntegreerd over het segment) de hoogste waarde bezit. Men bepaalt dan bijvoorbeeld:

$$O_x = \frac{1}{i_{\ell+1} - i_\ell} \int_{i_\ell}^{i_{\ell+1}} c_{i,\,px}\, dx$$

waarbij i_ℓ en $i_{\ell+1}$ de onder- en bovengrenzen van het segment zijn. Dat foneem x wordt nu toegekend aan het bewuste segment waarvoor geldt dat O_x de grootste waarde heeft.

Past men deze procedure op elk segment toe, dan vindt men uiteindelijk de rij van aaneengeschakelde fonemen waaruit de spraakuiting is opgebouwd. De spraakuiting is daarmee herkend.

Het zij vermeld dat de uitvinding niet beperkt is tot enkel de getoonde uitvoeringsvoorbeelden. Verschillende modifikaties van de beschreven uitvoeringsvoorbeelden zijn mogelijk zonder dat wordt afgeweken van de uitvinding zoals die wordt gedefinieerd door de konklusies.

BAD ORIGINAL

Tabel I - Omzettingsregels van foneemgrenzen naar difoongrenzen voor de belangrijkste fonemen van de Nederlandse taal.

Foneem-nummer	Foneem	Voorbeeld	Difoongrenzen
1	I	pit	5 intervallen na begin foneem
2	i	liep	6 intervallen na begin foneem
3	ɛ	les	6 intervallen na begin foneem
4	e	lees	6 intervallen na begin foneem
5	ɑ	mat	6 intervallen na begin foneem
6	ɑ	maat	10 intervallen na begin foneem
7	ɔ	rot	6 intervallen na begin foneem
8	o	rood	6 intervallen na begin foneem
9	u	roet	5 intervallen na begin foneem
10	y	fuut	5 intervallen na begin foneem
11	œ	put	5 intervallen na begin foneem
12	ə	de	6 intervallen na begin foneem
13	ɑu	koud	8 intervallen na begin foneem
14	ʌy	muis	8 intervallen na begin foneem
15	ø	keus	10 intervallen na begin foneem
16	eI	reis	8 intervallen na begin foneem
17	#	stilte	midden van foneem
18	ʔ	glottal stop	midden van foneem
19	p	pas	1 interval voor burst
20	t	tas	1 interval voor burst
21	k	kas	1 interval voor burst
22	b	bas	1 interval voor burst
23	d	das	1 interval voor burst
24	g	goal	1 interval voor burst
25	s	sok	midden van foneem
26	f	fok	midden van foneem
27	x	gok	midden van foneem
28	z	zeer	midden van foneem
29	v	veer	midden van foneem
30	m	meer	midden van foneem
31	n	neer	midden van foneem
32	ŋ	bang	midden van foneem
33	l	lang	midden van foneem
34	r	rang	midden van foneem
35	w	wang	midden van foneem
36	j	jan	midden van foneem

BAD ORIGINAL

Tabel II - Omzettingsregels van foneemgrenzen naar difoongrenzen voor de belangrijkste fonemen van de Duitse taal.

Phonem-nummer	Phonem	Beispiel	Diphongrenze
1	I	ich	4 Intervalle nach Beginn des Phonems
2	i	ihre	7 Intervalle nach Beginn des Phonems
3	ɛ	elf	4 Intervalle nach Beginn des Phonems
4	e	eben	7 Intervalle nach Beginn des Phonems
5	a	April	5 Intervalle nach Beginn des Phonems
6	ɑ	aber	10 Intervalle nach Beginn des Phonems
7	ə	Stunde	3 Intervalle nach Beginn des Phonems
8	ɔ	Onkel	4 Intervalle nach Beginn des Phonems
9	o	oder	8 Intervalle nach Beginn des Phonems
10	u	und	4 Intervalle nach Beginn des Phonems
11	ʊ	Uhr	7 Intervalle nach Beginn des Phonems
12	y	Büro	4 Intervalle nach Beginn des Phonems
13	ʏ	übel	8 Intervalle nach Beginn des Phonems
14	oe	öffentlich	5 Intervalle nach Beginn des Phonems
15	ø	Öfen	9 Intervalle nach Beginn des Phonems
16	ɐ	eifrig	5 Intervalle nach Beginn des Phonems
17	#	Stille	Mitte des Phonems
18	@	glottal stop	Mitte des Phonems
19	p	Pater	1 Intervall vor "Burst"
20	t	Tat	1 Intervall vor "Burst"
21	k	Kater	1 Intervall vor "Burst"
22	b	Bad	1 Intervall vor "Burst"
23	d	da	1 Intervall vor "Burst"
24	g	Gabel	1 Intervall vor "Burst"
25	s	das	Mitte des Phonems

BAD ORIGINAL

	26	f	Fahrrad	Mitte des Phonems
	27	x	acht	Mitte des Phonems
	28	s	sehen	Mitte des Phonems
	29	ʃ	schon	Mitte des Phonems
5	30	m	Mutter	Mitte des Phonems
	31	n	Natur	Mitte des Phonems
	32	ŋ	Junge	Mitte des Phonems
	33	l	lang	Mitte des Phonems
	34	r	rot	Mitte des Phonems
10	35	w	wahr	Mitte des Phonems
	36	j	ja	Mitte des Phonems
	37	h	hoffen	Mitte des Phonems
	38	ɔy	eure	5 Intervalle nach Beginn des Phonems
15	39	aʊ	auch	5 Intervalle nach Beginn des Phonems
	40	ç	ich	Mitte des Phonems

20

25

30

35

BAD ORIGINAL

5500377

PHN 11.275
Tabel III - Omzettingsregels van foneemgrenzen naar difoongrenzen voor de belangrijkste fonemen van de Engelse taal.

Phoneme number	Phoneme	Example	diphone boundary
1	I	bid	2 intervals after beginning of phoneme
2	i:	bead	7 intervals after beginning of phoneme
3	ɛ	bed	2 intervals after beginning of phoneme
4	ʌ	bud	2 intervals after beginning of phoneme
5	ei	bayed	9 intervals after beginning of phoneme
6	ɑi	bide	9 intervals after beginning of phoneme
7	ɔ	hot	2 intervals after beginning of phoneme
8	ɔi	foil	9 intervals after beginning of phoneme
9	u	hood	2 intervals after beginning of phoneme
10	ɑu	bowed	9 intervals after beginning of phoneme
11	iə	beard	9 intervals after beginning of phoneme
12	ə	sitter	2 intervals after beginning of phoneme
13	ɛə	bared	9 intervals after beginning of phoneme
14	uə	poor	9 intervals after beginning of phoneme
15	ɑe	bad	3 intervals after beginning of phoneme
16	ə:	bird	7 intervals after beginning of phoneme
17	#	silence	middle of the phoneme
18	@	glottal stop	middle of the phoneme
19	p	pin	1 interval before burst
20	t	tin	1 interval before burst
21	k	kin	1 interval before burst
22	b	bin	1 interval before burst
23	d	din	1 interval before burst
24	g	gap	1 interval before burst
25	s	sin	middle of the phoneme

BAD ORIGINAL

26	f	fin	middle of the phoneme
27	ʒ	pleasure	middle of the phoneme
28	z	zip	middle of the phoneme
29	ʃ	shin	middle of the phoneme
30	m	map	middle of the phoneme
31	n	nap	middle of the phoneme
32	ŋ	hang	middle of the phoneme
33	l	lap	middle of the phoneme
34	r	rap	middle of the phoneme
35	w	win	middle of the phoneme
36	j	yes	middle of the phoneme
37	h	hip	middle of the phoneme
38	θ	thin	middle of the phoneme
39	ð	this	middle of the phoneme

BAD ORIGINAL

8500377

Referenties

(1) B.A.G. Elsendoorn, J. 't Hart: Exploring the possibilities of speech synthesis with Dutch diphones.
IPO Annual Progress Report no. 17, 1982, p. 63-5.

(2) M. Stella: Fabrication semi-automatique de dictionnaires de diphones.
Recherche/Acoustique, Centre National d'Etudes de Télécommunication, vol VII, 1982/83, p. 53-63.

(3) L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon, T.M. Zampini:
A bootstrapping training technique for obtaining demisyllable reference patterns.
J. Acoust. Soc. America, 71(6), June 1982, p. 1588-95.

(4) B.S. Atal: Efficient coding of LPC-parameters by temporal decomposition.
Proc. of the ICASSP '83 (Boston), p. 81-4.

(5) J.D. Markel, A.H. Gray, Linear prediction of speech.
Springer Verlag 1976.

(6) L.L.M. Vogten: Analyse, zuinige codering en resynthese van spraakgeluid.
Thesis, Technische Hogeschool Eindhoven 1983.

(7) MEA 8000 Voice synthesizer, principles and interfacing.
Philips' Elcoma publikatie no. 101.

(8) R. Bellman: Dynamic programming
University Press, Princeton New Jersey, 1957.

(9) H. Ney: Dynamic programming as a technique for pattern recognition
Proc. of the 6th Int. Conf. on Pattern recognition, Munich (DE), Oct. 19-22, 1982, p. 1119-25.

BAD ORIGINAL

Konklusies:

1. Werkwijze voor het segmenteren van spraak, waarbij een akoestisch spraaksignaal wordt omgezet in N signalen, elk signaal behorende bij een tijdinterval i van N opvolgende tijdintervallen, waarbij i looptvan 1 tot en met N, met het kenmerk,

- dat voor opvolgende tijdintervallen i een funktie c_{ij} wordt afgeleid die een maat is voor de overeenkomst tussen het signaal behorende bij het tijdinterval i en het signaal behorende bij het tijdinterval j, waarbij voor een zekere i j loopt van a_i ($a_i < i$) tot en met b_i ($b_i > i$), en
- dat die waarden i, behorend bij een rangnummer i van een tijdinterval, worden bepaald die de begrenzingen vormen van gedeelten van het spraaksignaal waarvoor geldt dat de kurves c_{ij} behorend bij tijdintervallen liggend binnen een gedeelte ten minste ongeveer overeenkomen.

2. Werkwijze volgens konklusie 1, met het kenmerk,

- dat voor elk tijdinterval i het midden m_i van de bij het interval i behorende funktie c_{ij} wordt bepaald, en
- dat voor de genoemde waarden i die waarden i genomen worden die ten minste ongeveer overeenkomen met nuldoorgangen met eenzelfde teken in een funktie d_i, aangevende het verschil tussen het midden i en het rangnummer van het betreffende interval i.

3. Werkwijze volgens konklusie 2, met het kenmerk, dat die waarden i worden bepaald die ten minste ongeveer overeenkomen met nuldoorgangen van negatief naar positief in de funktie d_i, waarbij $d_i = m_i - i$.

4. Werkwijze volgens konklusie 2 of 3, met het kenmerk, dat a_i en b_i gelijk genomen worden aan die waarden i waarvoor geldt dat de funktie c_{ij} gelijk is aan een referentiewaarde c_r.

5. Werkwijze volgens konklusie 4, met het kenmerk, dat de referentiewaarde gelijk aan p maal de maximale waarde van de funktie c_{ij} genomen wordt, waarbij $p < 1$, bij voorkeur 0.8.

6. Werkwijze volgens konklusie 2 of 3, met het kenmerk, dat voor de N signalen de frekwentiespektra van het akoestisch spraaksignaal in de N tijdintervallen wordt genomen.

7. Werkwijze volgens konklusie 2 of 3, met het kenmerk, dat voor de N signalen de spektrale omhullenden van het akoestische spraaksignaal in de N tijdintervallen wordt genomen.

8. Werkwijze volgens konklusie 1, 2 of 3, voor het segmenteren van een spraaksignaal van een op zich bekende spraakuiting in segmenten van klanken, bijvoorbeeld in de vorm van fonemen, met het kenmerk, dat

BAD ORIGINAL

door middel van een op zich bekende tijdaanpassing tussen een aaneenschakeling van referentiepatronen van die klanken waaruit de spraakuiting is opgebouwd en het spraaksignaal van de spraakuiting een mate van overeenstemming wordt gerealiseerd die maximaal is, en dat die klanken aan die segmenten, begrensd door de waarden i, worden toegekend waarvoor de grenzen van klank en overeenkomstig segment de beste overeenstemming vertonen.

9. Werkwijze volgens konklusie 8, met het kenmerk, dat, daar waar één klank is toegekend aan twee naast elkaar liggende segmenten, de waarde i voor de grens tussen deze twee segmenten wordt weggelaten en dat, daar waar twee klanken zijn toegekend aan één segment, de grens tussen de twee klanken wordt toegevoegd aan de groep van waarden i.

10. Werkwijze volgens konklusie 9, met het kenmerk, dat uitgaande van de aldus verkregen groep van waarden i, gebruikmakend van omzettingsregels, de difoongrenzen worden afgeleid.

11. Werkwijze volgens konklusie 1, 2 of 3, met het kenmerk, dat voor het herkennen van het spraaksignaal van een onbekende spraakuiting, elk segment wordt vergeleken met referentiepatronen van op zich bekende klanken, en dat die klank aan een segment wordt toegekend waarvoor de mate van overeenstemming tussen dit segment en de klank maximaal is.

12. Inrichting voor het uitvoeren van de werkwzije volgens één der voorgaande konklusies, met een ingangsklem voor het ontvangen van een elektrisch signaal dat overeenkomt met het akoestische spraaksignaal, een eerste eenheid met een ingang gekoppeld met de ingangsklem en een uitgang, welke eerste eenheid is ingericht voor het afleiden van de N signalen en voor het toevoeren van deze N signalen aan zijn uitgang, met het kenmerk, dat de inrichting verder is voorzien van:

- een tweede eenheid met een ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke tweede eenheid is ingericht voor het afleiden van de funktie c_{ij} voor opvolgende intervallen i uit de signalen behorende bij de tijdintervallen i en j en voor het toevoeren van de funktie c_{ij} aan zijn uitgang, en
- een deelinrichting met een ingang gekoppeld met de uitgang van de tweede eenheid en een uitgang, welke deelinrichting is ingericht voor het bepalen van die waarden i die de begrenzingen vormen van gedeelten van het spraaksignaal waarvoor geldt dat de kurves c_{ij} behorend bij tijdintervallen liggend binnen een gedeelte ten minste ongeveer overeenkomen.

BAD ORIGINAL

13. Inrichting volgens konklusie 12, voor het uitvoeren van de werkwijze volgens konklusie 2, met het kenmerk, dat de deelinrichting is voorzien van:

- een derde eenheid met een ingang gekoppeld met de ingang van de deelinrichting en een uitgang, welke derde eenheid is ingericht voor het voor elk interval i afleiden van het midden m_i van de bijbehorende funktie c_{ij} en voor het toevoeren van de waarden m_i aan zijn uitgang, en
- een vierde eenheid met een ingang gekoppeld met de uitgang van de derde eenheid en een uitgang gekoppeld met de uitgang van de deelinrichting, welke vierde eenheid is ingericht voor het bepalen van die waarden i die ten minste ongeveer overeenkomen met nuldoorgangen in de funktie d_i en voor het toevoeren van die waarden i aan zijn uitgang.

14. Inrichting volgens konklusie 13, voor het uitvoeren van de werkwijze volgens konklusie 3, met het kenmerk, dat vierde eenheid is ingericht voor het bepalen van die waarden i die ten minste ongeveer overeenkomen met de nuldoorgangen van negatief naar positief in de funktie d_i, waarbij $d_i = m_i - i$.

15. Inrichting volgens konklusie 13, voor het uitvoeren van de werkwijze volgens konklusie 7, met het kenmerk, dat de eerste eenheid is ingericht voor het afleiden van de N signalen in de vorm van de spektrale omhullenden van het akoestische spraaksignaal in de N tijdintervallen.

16. Inrichting volgens één der konklusies 13, 14 of 15, voor het uitvoeren van de werkwijze volgens konklusie 8, met het kenmerk, dat de inrichting verder is voorzien van een geheugen voor het opslaan van de referentiepatronen van klanken, een vijfde eenheid met een eerste ingang gekoppeld met een uitgang van het geheugen, een tweede ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke vijfde eenheid is ingericht voor het door middel van een op zich bekende tijdaanpassing tussen een aaneenschakeling van referentiepatronen van die klanken waaruit de op zich bekende spraakuiting is opgebouwd en het spraaksignaal van de spraakuiting, realiseren van een maximale mate van overeenstemming en voor het leveren van de op deze wijze verkregen grenzen aan zijn uitgang, en een kombinatie-eenheid met een eerste ingang gekoppeld met de uitgang van de vierde eenheid en een tweede ingang gekoppeld met de uitgang van de vijfde eenheid en een uitgang, welke vijfde eenheid is ingericht voor het toekennen van die klanken aan die segmenten begrensd door de waarden i, waarvoor de grenzen van klank en overeenkomstig segment de beste overeen-

BAD ORIGINAL

stemming vertonen.

17. Inrichting volgens konklusie 16, met het kenmerk, dat de kombinatie-eenheid verder is ingericht voor het toevoegen van en het weglaten van één of meer waarden i en wel net zoveel waarden i als er minder respektievelijk meer segmenten zijn dan overeenkomstige klanken en voor het leveren van de al dan niet veranderde groep van waarden i aan zijn uitgang.

18. Inrichting volgens konklusie 17, met het kenmerk, dat de inrichting verder is voorzien van een zesde eenheid, met een ingang gekoppeld met de uitgang van de kombinatie-eenheid en een uitgang, welke zesde eenheid is ingericht voor het, uitgaande van de waarden i van de kombinatie-eenheid voor de klankgrenzen en gebruikmakend van omzettingsregels, afleiden van de difoongrenzen en voor het toevoeren van deze difoongrenzen aan zijn uitgang.

19. Inrichting volgens konklusie 13 of 14, voor het uitvoeren van de werkwijze volgens konklusie 11, met het kenmerk, dat de inrichting verder is voorzien van een geheugen voor het opslaan van referentiepatronen van op zich bekende klanken en van een vijfde eenheid met een eerste ingang gekoppeld met de uitgang van het geheugen, een tweede ingang gekoppeld met de uitgang van de vierde eenheid en een derde ingang gekoppeld met de uitgang van de eerste eenheid en een uitgang, welke vijfde eenheid is ingericht voor het vergelijken van elk segment met referentiepatronen van op zich bekende klanken, en voor het toekennen van die klank aan een segment waarvoor de mate van overeenstemming tussen dit segment en de klank maximaal is, en is ingericht voor het leveren aan zijn uitgang van informatie betreffende de aaneenschakeling van referentiepatronen van die klanken die elk maximaal overeenstemmen met het segment waaraan zij zijn toegekend.

BAD ORIGINAL

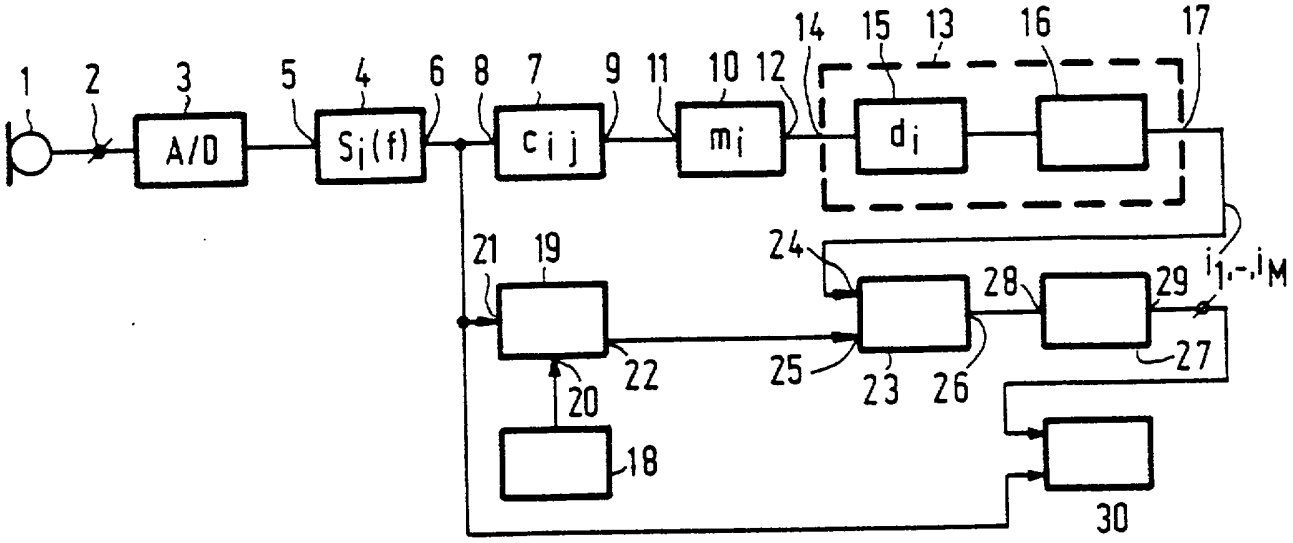


FIG. 1

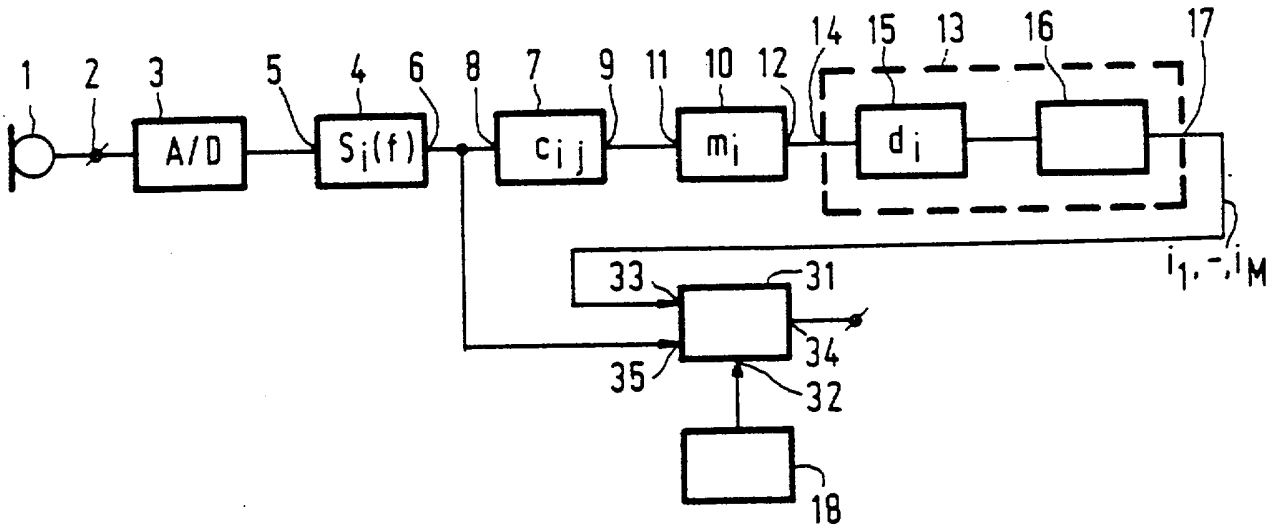


FIG. 2

BAD ORIGINAL

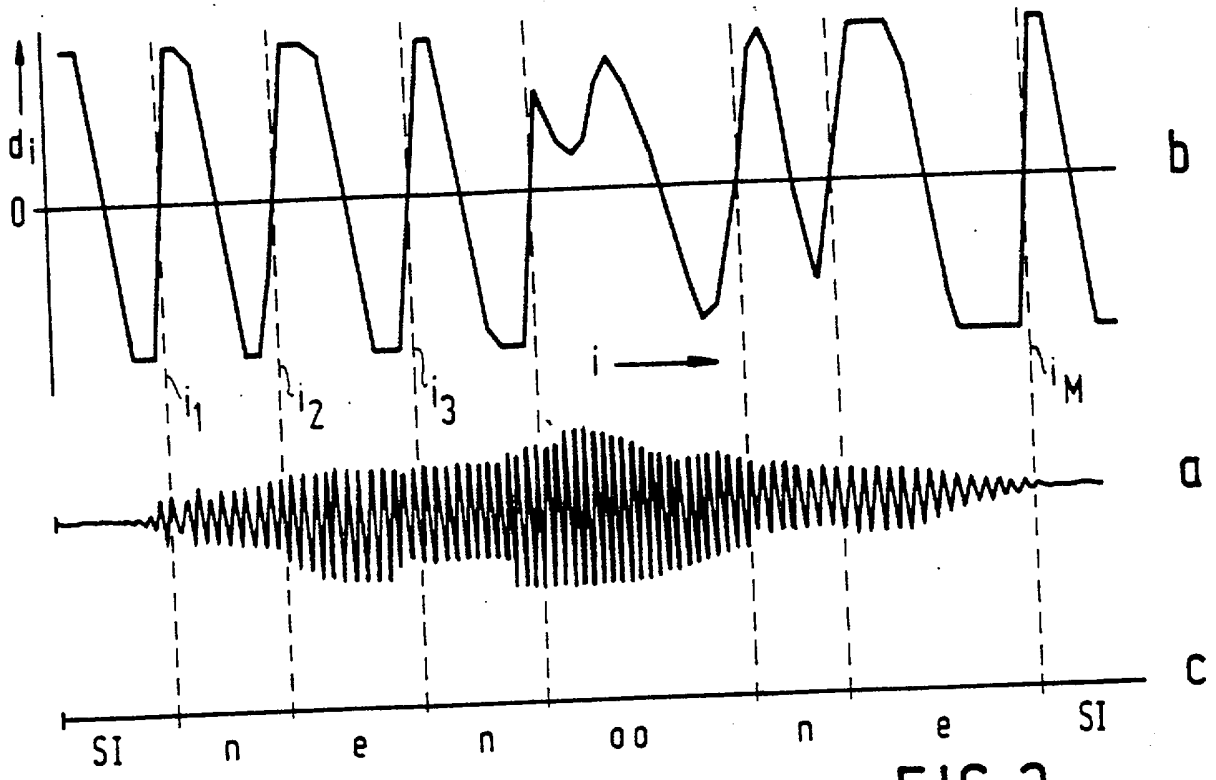
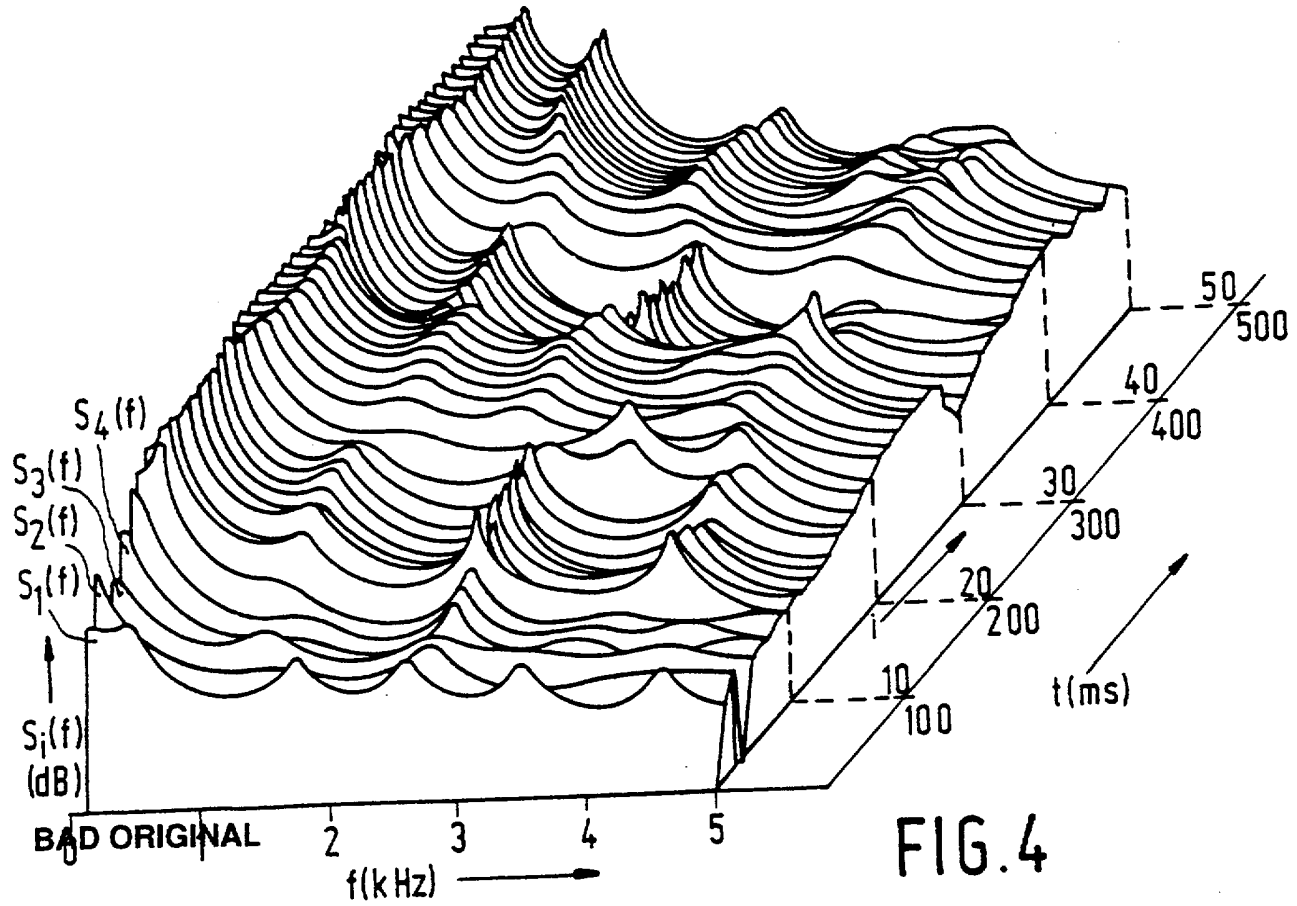


FIG.3



BAD ORIGINAL

FIG.4

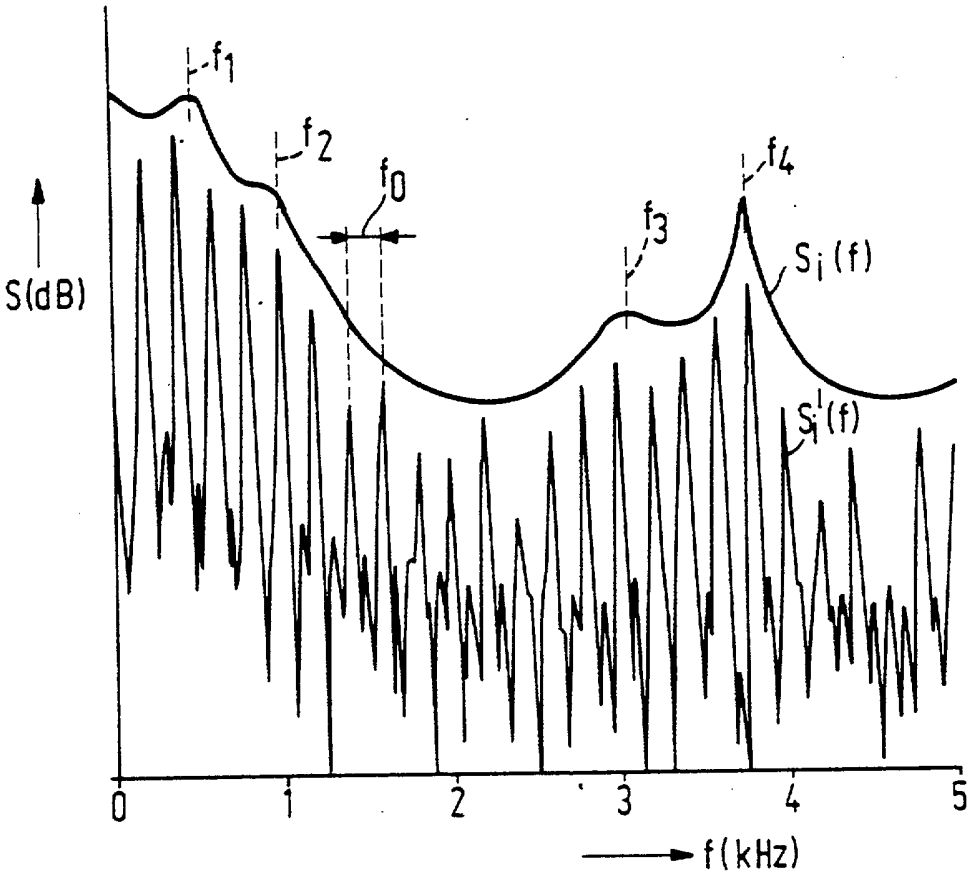


FIG.5

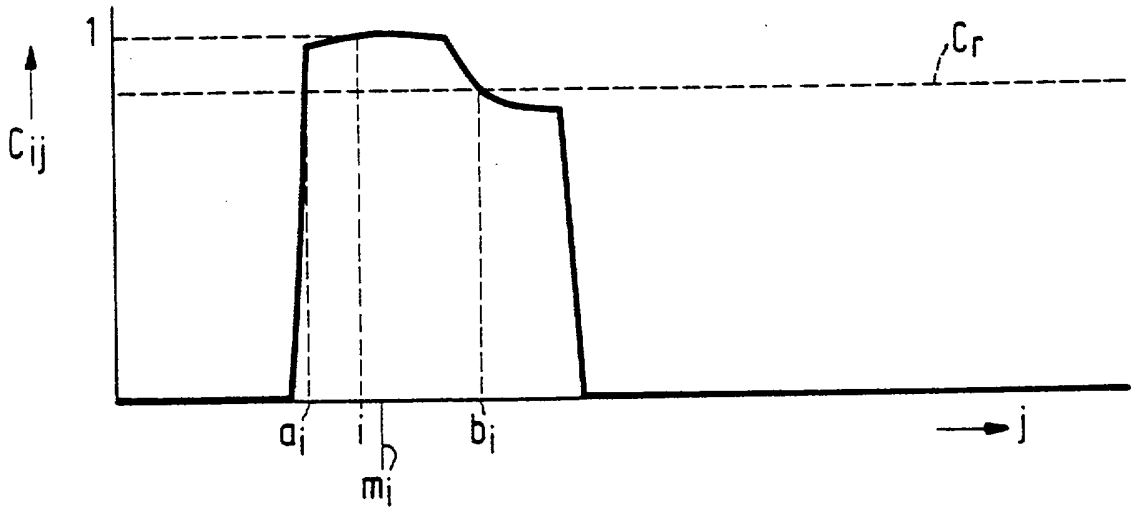


FIG.6

BAD ORIGINAL

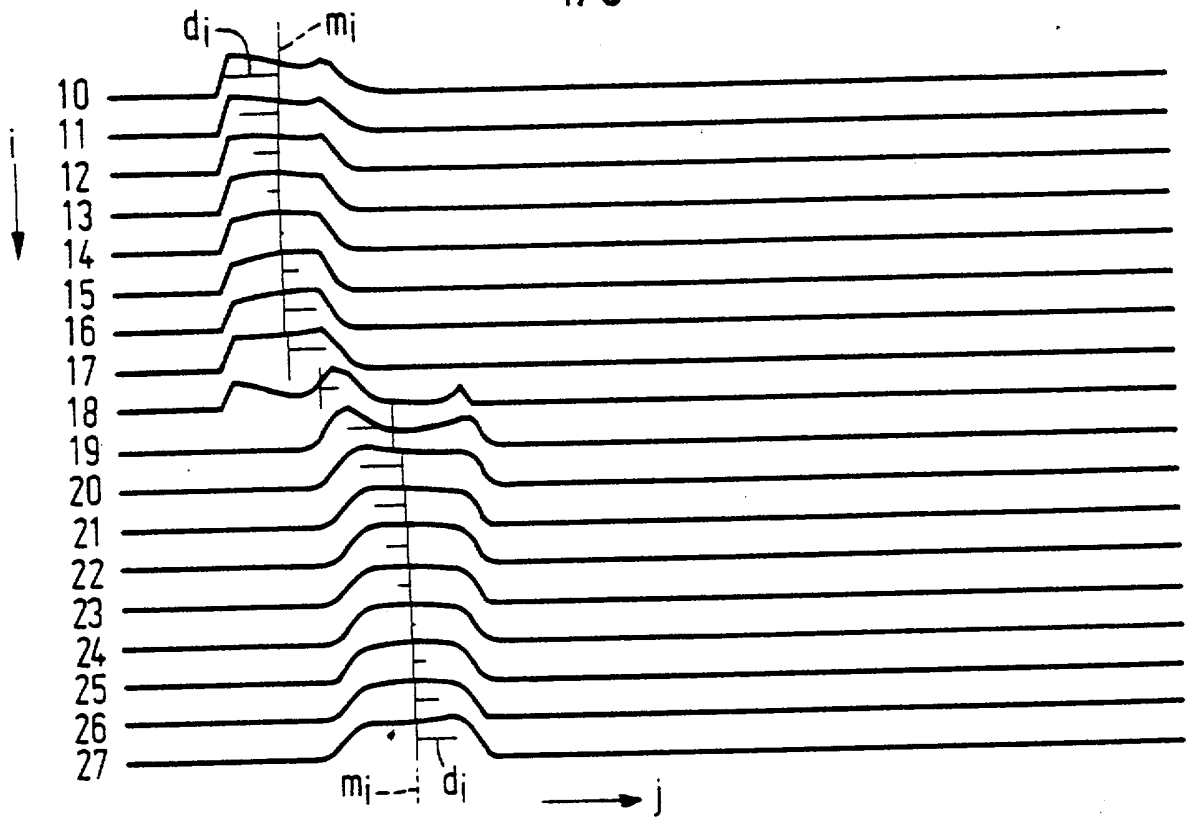

di
mi
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
i
mi
di
j

FIG.7

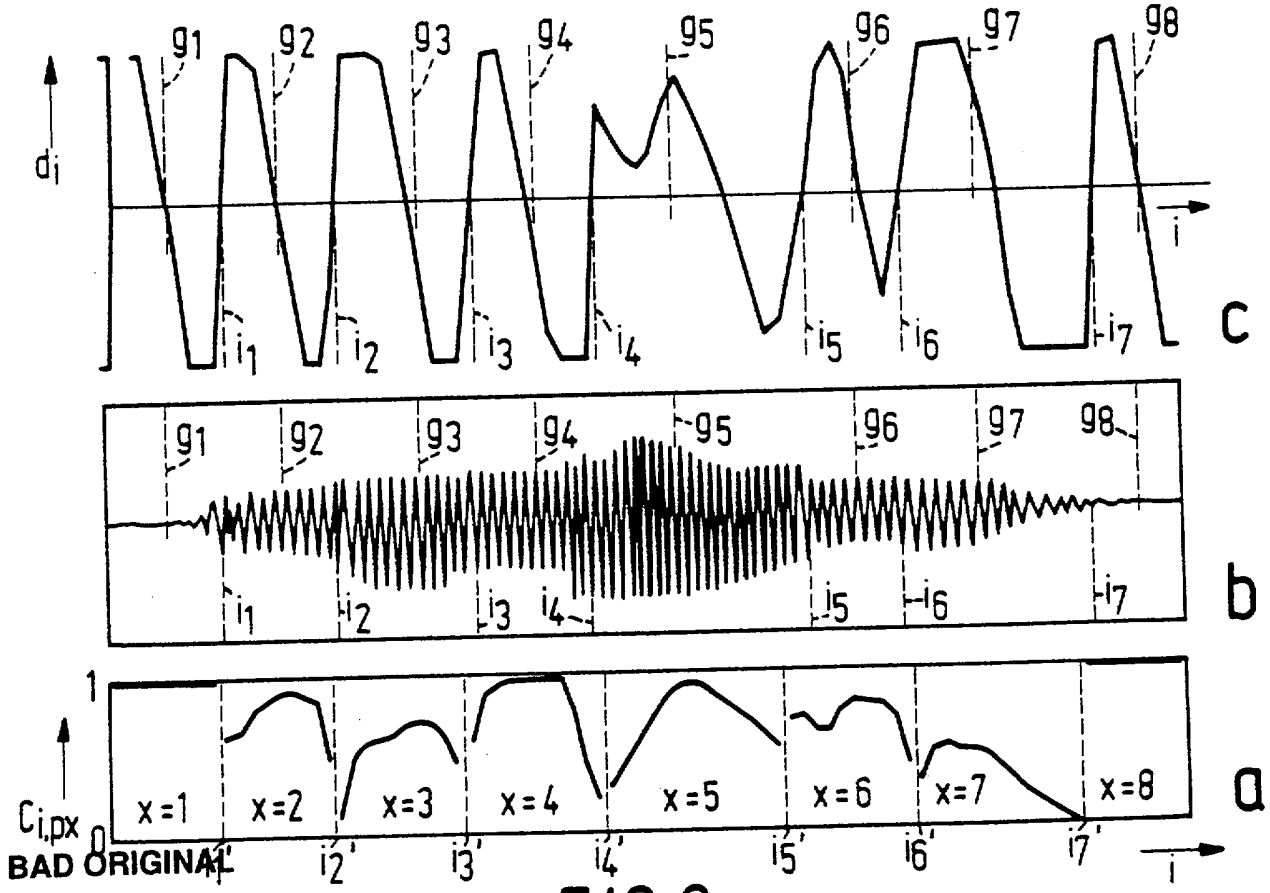

di
g1
g2
g3
g4
g5
g6
g7
g8
i1
i2
i3
i4
i5
i6
i7
i
c
b
1
0
Ci,px
x=1
x=2
x=3
x=4
x=5
x=6
x=7
x=8
i1'
i2'
i3'
i4'
i5'
i6'
i7'
i
a
BAD ORIGINAL

FIG.8

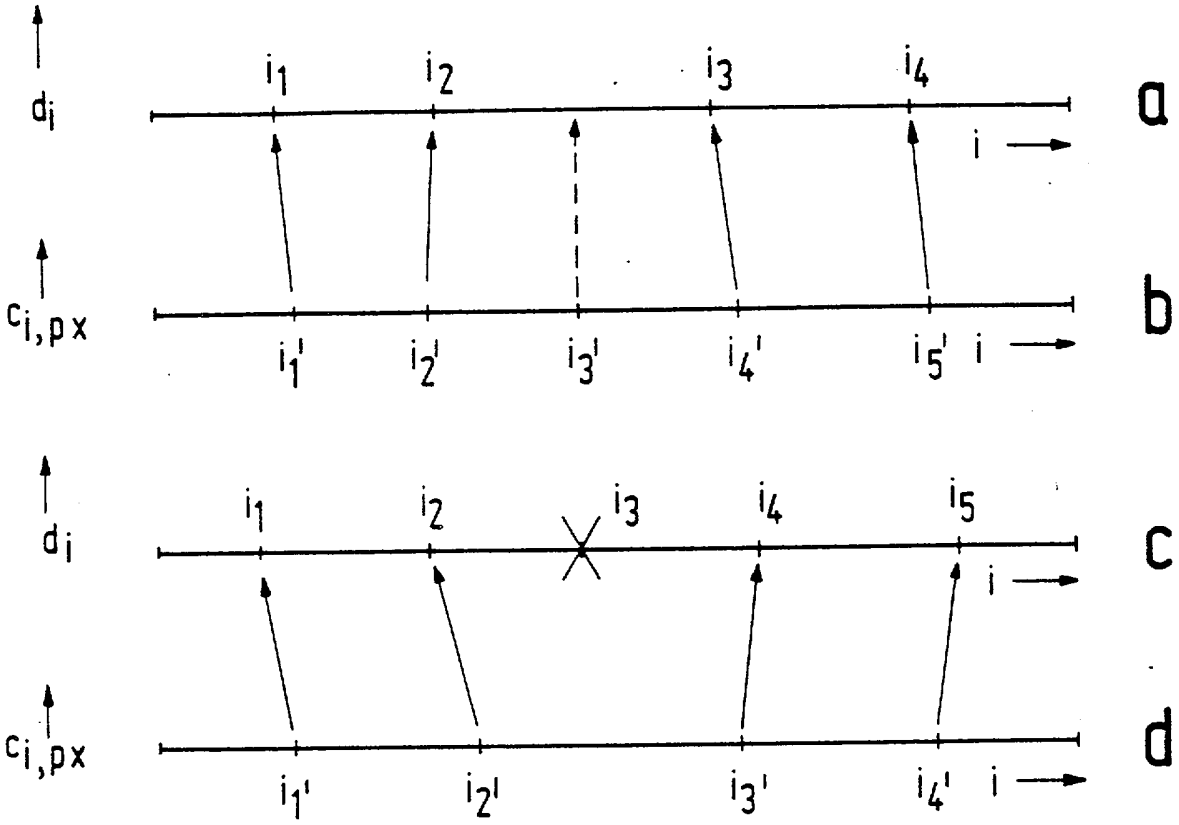


FIG.9

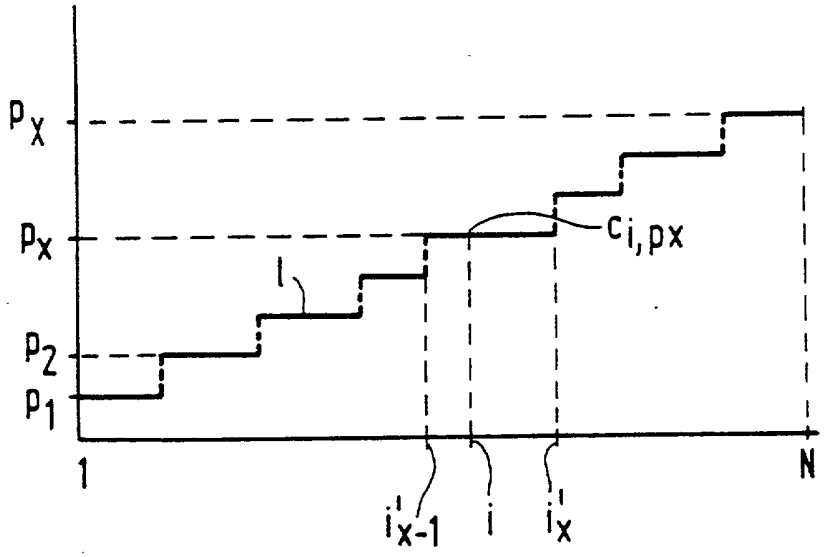


FIG.10

BAD ORIGINAL