Spraakonderzoek in het instituut voor Perceptie Onderzoek

P376 TH Berichten 22/1980, no. 31. 18-04-80

Spraakonderzoek in het Instituut voor Perceptie Onderzoek

In veel situaties, waarin feitelijke, zakelijke informatie tussen mensen moet worden overgebracht, is spraak ongetwijfeld het snelste, het meest natuurlijke en het meest flexibele medium. Sinds de uitvinding van de telefoon een einde maakte aan de beperkte reikwijdte van de menselijke stem, is het belang van spraak voor de uitwisseling van informatie tussen mensen onderling nog toegenomen. Daarnaast vindt in onze moderne samenleving in toenemende mate informatieverkeer plaats tussen mensen en meer of minder ingewikkelde apparaten. Gezien de voor de mens plezierige eigenschappen van spraak als communicatiemiddel ligt het voor de hand na te denken over de mogelijkheden om informatie-uitwisseling tussen mensen en apparaten geheel of gedeeltelijk via spraak te laten lopen.
In heel beperkte mate is dit het geval bijvoorbeeld bij de automatische tijdmelding, weerberichten etc. Ook zijn er al praktische toepassingen van (eenvoudige) systemen voor de herkenning van geïsoleerd gesproken woorden uit een beperkt vocabulair, gesproken door één spreker. Meer ingewikkelde vormen van informatie-uitwisseling tussen mens en apparatuur via spraak leken tot voor enkele jaren nog heel ver in de toekomst te liggen.
Door de snelle ontwikkelingen in elektronische technieken, met name het steeds sneller en goedkoper worden van digitale processoren en geheugens, is er een verschuiving gekomen in deze toekomstverwachting. Wat vroeger heel ver weg leek, komt nu veel sneller dichtbij dan men verwacht had. Het is op het ogenblik te verwachten, dat binnen een aantal jaren de spraaktechnologie tot belangrijke toepassingen zal leiden op het gebied van de informatie-uitwisseling tussen mensen en hun apparatuurlijke omgeving. Men kan daarbij denken aan het toespreken van en toegesproken worden door computers in geautomatiseerde informatiesystemen en in systemen voor het besturen van apparatuur in geautomatiseerde onderwijssystemen en aan hulpmiddelen voor gehandicapten, zoals 'leesmachines' voor blinden, sprekende typemachines enz.
Voor een zinvolle verwezenlijking van deze mogelijkheden is veel fundamenteel onderzoek nodig. Zulk onderzoek zal vooral gericht zijn op het doorzichtig maken van menselijke communicatie door middel van spraak. Een belangrijk deel van dit onderzoek betreft het leren begrijpen van het verband tussen de fysische eigenschappen van spraakgeluid aan de ene kant en de communicatieve functies van spraak aan de andere kant.
Bij het onderzoek naar de communicatieve functies van spraak, is het noodzakelijk de fysische eigenschappen van spraak te beheersen, zodat deze als experimentele variabelen in luisterexperimenten gebruikt kunnen worden.

Al sinds kort na de oprichting van het Instituut voor Perceptie Onderzoek (IPO) maakt spraakonderzoek deel uit van het onderzoeksprogramma. Over dit spraakonderzoek, de gebruikte methodiek en enkele resultaten ervan, handelt deze bijdrage. Eerst vertellen we iets over de fysische structuur van spraakgeluid en een daarmee samenhangend spraakproduktiemiddel. Dan geven we een korte beschrijving van een op dit model gebaseerd analyse-resynthese systeem zoals dat bij het spraakonderzoek gebruikt wordt en we besluiten met enkele voorbeelden van onderzoeksresultaten en hun praktische toepassingen.

L.L.M. Vogten

De fysische structuur van spraak: spraakproduktiemodel
Om een indruk te geven van de ingewikkeldheid van spraakgeluid laten we in fig. 1 een registratie zien van de golfvorm van een spraakuiting, zoals die bv. via een microfoon op magneetband is opgenomen. We kunnen daaraan een aantal verschijnselen opmerken.

We zien bijvoorbeeld midden in de spraakuiting korte geluidloze segmenten, die meestal samengaan met 'plofklanken' zoals p, t en k. Verder zien we dat quasi-periodieke golfvormen, 'klinkers', afwisselen met niet-periodieke. De gemiddelde amplitude vertoont een grillig verloop over de hele uiting. Duren van akoestische segmenten, die bij benadering corresponderen met individuele spraakklanken variëren sterk.
Inspectie van dergelijke registraties van de golfvorm van spraakuitingen leert ons wel, dat spraakgeluid ingewikkeld is, maar leidt niet onmiddellijk tot een analyse van spraaksignalen in termen van geschikte parameters. Om tot zo'n analyse te komen is het zinvol gebleken om spraak op te vatten als het resultaat van een of meer brongeluiden, die gefilterd worden door een akoestisch filter dat wordt gevormd door de mond-keelholte (en bij sommige spraakklanken door de neusholte), zoals in fig. 2 is weergegeven.

We onderscheiden in het algemeen twee typen bron geluiden:

a. een quasi-periodieke golfvorm, rijk aan boventonen, teweeggebracht door het trillen van de stembanden. Dit is het brongeluid van alle klinkers en een aantal zogenaamde 'stemhebbende' medeklinkers, zoals m, n, r, l in niet gefluisterde spraak.

De herhalingsfrequentie van deze golfvorm is grotendeels *onder controle* van de spreker en bepaalt de toonhoogte (zinsmelodie) die de luisteraar waarneemt;

b. een ruisachtig brongeluid, veroorzaakt door turbulentie van de luchtstroom uit de longen in een vernauwing van het mondkanaal, bijvoorbeeld tussen tong en verhemelte of tussen boventanden en onderlip. Dit is het brongeluid van zogenaamde 'stemloze wrijfklanken' als f, s, ch. Bij 'stemhebbende' wrijfklanken als v en z gaat zo'n ruisgeluid, meestal aanzienlijk zwakker dan bij stemloze, samen met de periodieke golfvorm afkomstig van de stembandtrilling. Bij plofklanken zoals p, t, k wordt een kortdurend ruisgeluid voorafgegaan door een akoestische stilte, tengevolge van een volledige afsluiting in het mondkanaal. Wanneer deze afsluiting wordt opgeheven, ontstaan extra snelle overgangsverschijnselen aan het begin van de ruisstoot.

Beide brongeluiden hebben een breed of 'vlak' spectrum, d.w.z. beide zijn rijk aan zowel lage als hoge tonen. Door de overdrachtskarakteristiek van de keel-mondholte wordt dit vlakke spectrum gefilterd of 'gekleurd' tot de ons bekende spraakklanken. Hoe die overdrachtskarakteristiek eruit ziet, wordt in eerste benadering bepaald door de akoestische eigenschappen van de gezamenlijke holten, die boven de geluidsbron liggen.

Bij klinkergeluiden is dit de keel-mondholte, vanaf de stembanden in het strottehoofd tot aan de mondopening. Bij spraakklanken gevormd met een vernauwing in de mondholte is de werkzame resonantieruimte veel kleiner. Een mathematische methode die de akoestische eigenschappen van keel-, mond- en neusholte nauwkeurig beschrijft, is er niet. Wel zijn er verschillende manieren om de werkelijkheid zo te vereenvoudigen, dat een bruikbare

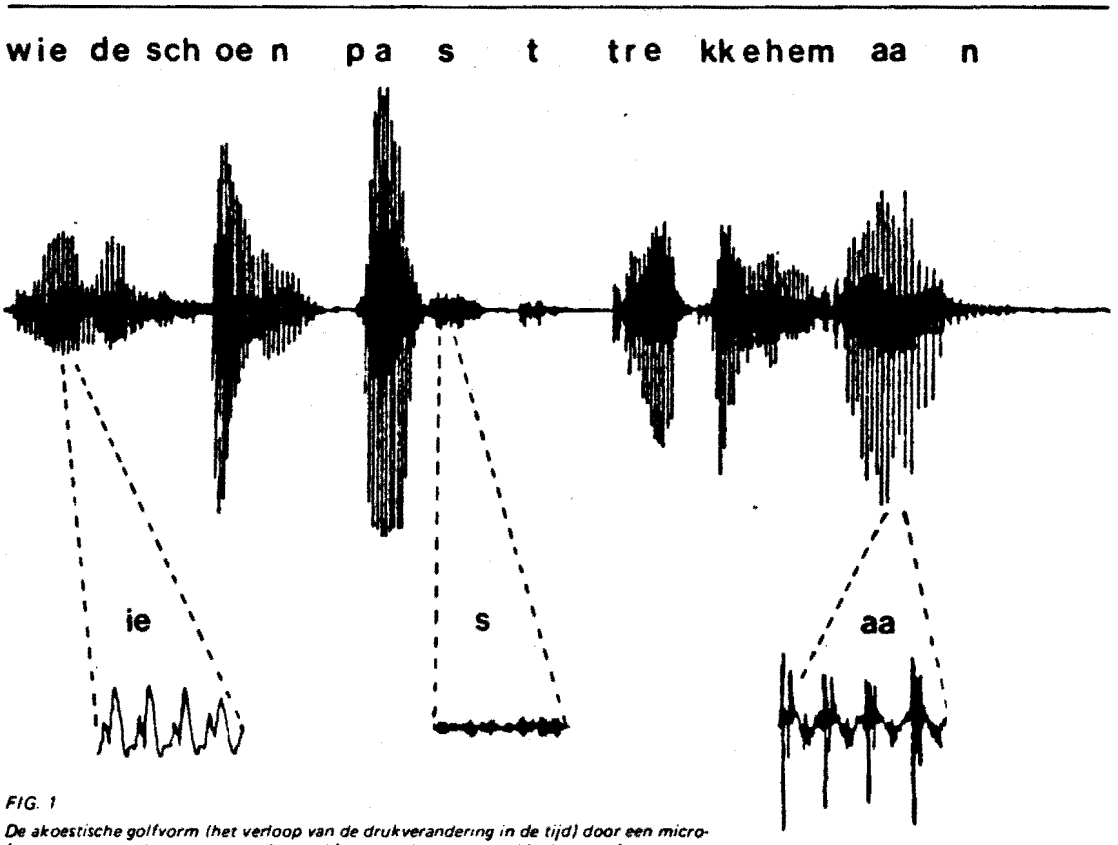

FIG. 1

De akoestische golfvorm (het verloop van de drukverandering in de tijd) door een microfoon geregistreerd voor het spreekwoord 'wie de schoen past trekke hem aan'. Details van twee klinkers en een stemloze wrijfklank zijn vergroot weergegeven, zodat het periodieke en het ruisachtige karakter van deze klanken te zien is.

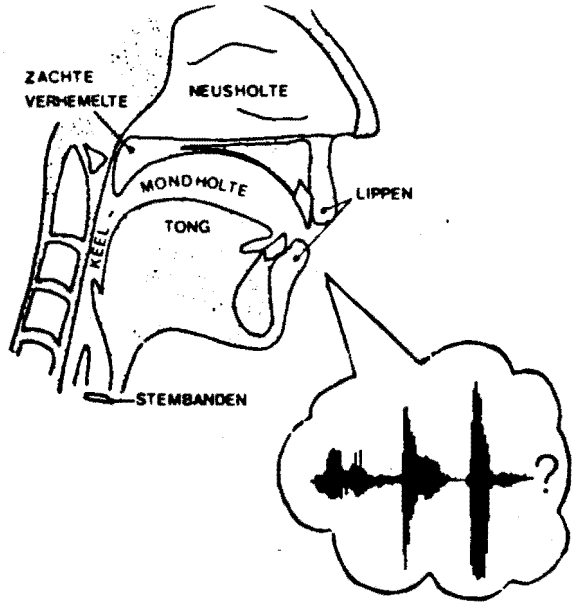

FIG. 2

Schematische weergave van het menselijk spraakmechanisme. Bij stemhebbende klanken brengen de stembanden en dit brongeluid wordt door resonanties in de keel-, neus- en mondholte 'gekleurd' tot spraakklanken. Voor stemloze klanken zijn de stembanden buiten werking en wordt het brongeluid gevormd door luchtturbulentie bij een plaatselijke vernauwing in de mond-keelholte. De voortdurend veranderende stand van kaken, tong, zachte verhemelte en lippen bepaalt welke opeenvolging van klanken wordt geproduceerd.

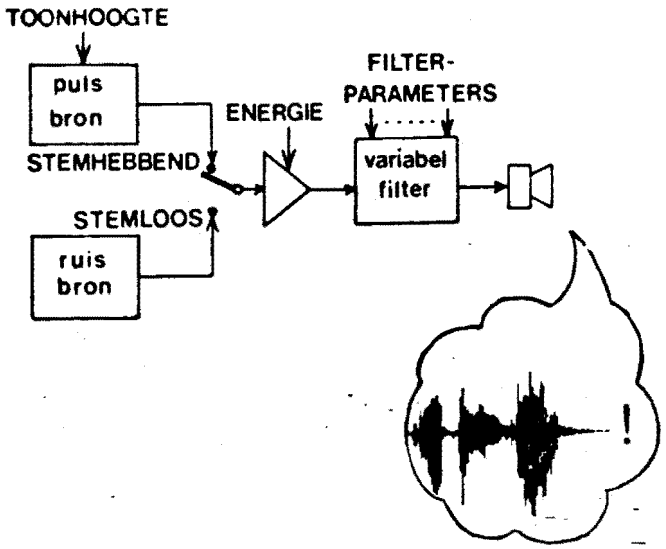

FIG. 3

Eenvoudig model voor de spraakproduktie. Een pulsgenerator produceert pulsen met variabele periodeduur en hoogte daarmee de stembandtrillingen na. Een ruisgenerator fungeert als bron voor stemloze klanken. Welke van de twee is ingeschakeld wordt bepaald door de stemhebbend/stemloos-stand. Na de regelbare versterker wordt het bronsignaal in het variable filter, dat de resonanties van de mond-keelholte nabootst, 'gekleurd' tot spraakklanken.

benadering kan worden verkregen.
Eén zo'n manier is het beschrijven van spraakgeluid met behulp van een eenvoudig lineair model, waarin de belangrijkste fysische grootheden zijn vertegenwoordigd, die een rol spelen bij de spraakproduktie.
Dat model, weergegeven in fig. 3, bestaat uit twee signaalbronnen - een pulsgenerator met variabele pulsherhalingsfrequentie voor de 'stemhebbende' klanken en een ruisgenerator voor de stemloze klanken, alsmede een variabel filter dat de overdrachtskarakteristiek van de keel-mondholte beschrijft.
Daarnaast bepaalt een schakelelement welke van de twee bronnen aan het filter wordt toegevoerd en een versterker regelt de energie (luidheid) van de te produceren klank. Tenslotte zal bij elektronische realisatie van het model het uitgangssignaal van het filter nog omgezet moeten worden in een geluidstrilling bv. via een luidspreker.
Met dit model is het mogelijk allerlei spraakklanken te genereren, van losse klinkers tot complete zinnen.
Hoe goed en hoe natuurlijk deze synthetische spraak dan klinkt hangt o.a. ervan af in hoeverre wij in staat zijn om de afzonderlijke parameters van het model op de juiste wijze en op het juiste tijdstip bij te stellen.
In eerste instantie is dat een hele moeilijke opgave; we weten immers niet hoe de afzonderlijke parameters in de tijd moeten verlopen om een samenhangende en vloeiende spraakuiting te verkrijgen. Er moet aan een groot aantal regels worden voldaan, die we nog maar zeer ten dele kennen en begrijpen.
Mede dankzij de al genoemde ontwikkeling van de digitale modelparameters snel rechtstreeks uit het spraaksignaal zelf te bepalen (spraakanalyse). Hoe we dat doen zal in de volgende paragraaf aan de orde komen.
Intussen zal wel duidelijk zijn, dat wanneer we in staat zijn om de modelparameters uit het spraaksignaal zelf te bepalen, we overeenkomstig het model de geanalyseerde spraak weer kunnen samenstellen tot de oorspronkelijke spraakuiting (spraakresynthese).
Daarmee hebben we dan een machtig stuk onderzoeksgereedschap in handen, waarmee we de afzonderlijke fysische grootheden kunnen beheersen, hun effect op de spraakwaarneming kunnen bestuderen en aldus kunnen onderzoeken welke van die fysische grootheden essentieel zijn voor de gehoorsindruk, welke bijdragen tot het verstaan en welke bijdragen tot het begrijpen van de taalboodschap.

Spraakanalyse en -resynthese

Om de gezochte parameters van het spraakproduktiemodel te bepalen maken we gebruik van het gegeven, dat het ingangssignaal voor het variabel filter een vlak spectrum heeft. Immers, als we een gegeven spraaksignaal zodanig filteren dat het spectrum ervan vlak wordt, hebben we daarvoor een filter nodig dat precies de omgekeerde versie is van het filter uit het produktiemodel. Dit levert ons nu de mogelijkheid om de modelparameters te bepalen voor een spraaksignaal dat eens door een menselijke stem is voortgebracht.
Zo'n spraakanalyse kan snel worden uitgevoerd met behulp van een digitale rekenmachine. Hiertoe wordt de geregistreerde spraakuiting omgezet in een rij getallen door de analoge golfvorm om de 100 microseconden te bemonsteren. Deze rij getallen wordt dan ingelezen in het schijvengeheugen van de rekenmachine. Bij een precisie van 12 bits per bemonstering kost ons dat voor één seconde spraak 120 000 bits aan geheugenruimte. Er zijn zuiniger manieren om gedigitaliseerde spraak op te slaan, maar daar gaan we hier niet op in. Hoofdzaak is dat we de spraakuiting in digitale vorm beschikbaar hebben om de analyse op uit te voeren. Deze gaat dan als volgt: we nemen een stukje spraak van 25 msec (250 bemonsteringen), berekenen daarvoor de 10 parameters van het digitale filter dat het spectrum zo vlak mogelijk maakt, en inverteren dit filter. Daarmee hebben we de gezochte filterparameters berekend voor dat stukje spraak waarbij de totale energie als 'bijprodukt' van de berekening beschikbaar komt.
Daarna bepalen we of het spraaksegment stemhebbend (quasi-periodiek) is of niet. Zo ja, dan berekenen we de periodeduur die de herhalingsfrequentie van de pulsbron (toonhoogte) weergeeft.
Vervolgens schuiven we 10 msec of 100 bemonsteringen op en berekenen opnieuw de filterparameters en energie, de stemhebbend/stemloos-stand en eventueel de toonhoogte.
Aldus doorlopen we de gehele spraakuiting in stapjes van 10 msec, waarbij ieder stapje een 'frame' van 13 parameters oplevert, die successievelijk worden opgeslagen in het schijvengeheugen van de rekenmachine. Daarmee is de eigenlijke spraakanalyse voltooid.
Een voorbeeld van zo'n analyseresultaat is weergegeven in fig. 4.
De totale duur van het spreekwoord is ongeveer 1,8 sec., overeenkomend met 180 frames van ieder 10 msec.
De bovenste helft van het plaatje laat zien hoe de parameters van het bronsignaal flucturen in de tijd; de toonhoogte veel geleidelijker dan de energie.
De onderste helft toont 5 van de 10 filterparameters die de mond-keelholte beschrijven.
Uit dit plaatje kunnen we opnieuw zien dat de fysische structuur van spraak, geanalyseerd in z'n beschrijvende parameters, nog steeds ingewikkeld is en veel grillige fluctuaties vertoont. Toch is voor de onderzoeker uit deze beschrijving al heel wat meer af te leiden dan uit de registratie van de oorspronkelijke golfvorm van fig. 1. Belangrijk is echter, dat de aldus geanalyseerde parameters gebruikt worden om de oorspronkelijke spraakuiting te reconstrueren. Deze resynthese gebeurt precies zoals het produktiemodel van fig. 3 aangeeft.
Iedere 10 ms worden toonhoogte (pulsherhalingsfrequentie), stemhebbend/stemloos-stand, energie en filterparameters bijgesteld overeenkomstig de gevonden analyseresultaten. Of deze resynthese nu in software met de rekenmachine dan wel in hardware met een speciaal daarvoor gebouwd apparaat wordt uitgevoerd, doet in principe niet terzake. In beide gevallen is de gereproduceerde spraak van goede kwaliteit, soms zelfs moeilijk te onderscheiden van de oorspronkelijke spraak.
Interessant voor het spraakonderzoek is de mogelijkheid om de afzonderlijke fysische parameters naar wens te veranderen door de getalwaarde van de analyse-uitkomst te wijzigen en het effect daarvan te beluisteren. Hiervan zal in de volgende paragraaf een aantal voorbeelden worden gegeven.

Onderzoeksresultaten en toepassingsvoorbeelden

Een eerste illustratie van de toepassing van het besproken analyse-resynthesesysteem is het gebruik ervan bij het IPO-intonatie-onderzoek. Alvorens te resynthetiseren kan de oorspronkelijke geanalyseerde toonhoogtecontour verwijderd worden door alles 'stemloos' te maken en alleen de ruis als bronsignaal te gebruiken. Dan ontstaat 'fluisterspraak'.
Ook kunnen alle variaties in de toonhoogte worden weggelaten door de herhalingsfrequentie van de pulsbron vast te zetten, waarmee volstrekt monotone spraak ontstaat.
Daarbij blijft de verstaanbaarheid vrijwel volledig intact. Waar het bij het intonatie-onderzoek echter om gaat is de vraag welke toonhoogteveranderingen wel en welke niet van belang zijn voor de waargenomen zinsmelodie. Daartoe kunnen we het toonhoogteverloop meer of minder sterk stileren en nagaan wanneer de luisteraar verschillen waarneemt.
Dan blijft er maar een klein aantal elementaire variaties over, die van belang blijken voor de melodische indruk. Vaak liggen die op plaatsen

vervolg zie pagina 8

In een dr zi aan de Academische Raad bevestigt [illegible] vrijstelling van betaling van collegegeld na vijf jaar te laten vervallen.
Op korte termijn zal hiertoe een wetsontwerp worden ingediend. Daarna dringt Pais aan op een snel advies van de Academische Raad en de Onderwijsraad.

Van de redactie

Pais motiveert z'n voornemen als volgt: 'Zeker in de huidige benarde situatie, waarin 's Rijksfinanciën verkeren, acht ik het uit een oogpunt van rechtvaardigheid en billijkheid redelijk, dat voor het gebruik maken van kostbare voorzieningen als het volgen van wetenschappelijk onderwijs een bijdrage waarin door de student van die voorzieningen gebruik wordt gemaakt.
De regeling vrijstelling van betaling van collegegeld na vijf jaar is bovendien niet consistent met het streven naar verkorting van de studieduur'.
De minister wil het wetsvoorstel snel indienen om een eventuele invoering in het komende studiejaar al mogelijk te maken.

Spraakonderzoek

vervolg van pagina 5

in de zin waar ook een accent ('klemtoon') ligt. Deze uitkomsten van het onderzoek zijn bijvoorbeeld van direct belang voor een efficiënt onderwijs in de uitspraak van het Nederlands.
Zo is er in het IPO een cursus Nederlandse intonatie samengesteld. Soortgelijk onderzoek wordt verricht aan de Engelse intonatie.
Een tweede voorbeeld waar het analyse-resynthesesysteem gebruikt wordt, is bij het onderzoek naar bewerkingen die nodig zijn om los ingesproken worden zó samen te voegen, dat een voor het oor vloeiend geheel ontstaat.
De genoemde kennis van de elementaire toonhoogtebewegingen is daarvoor essentieel, maar ook gedetailleerde kennis over de precieze opbouw in de tijd van de spraakuiting is onontbeerlijk om tot een acceptabel resultaat te komen.
Ook bij dit onderzoek is het analyse-resynthesesysteem een nuttig gereedschap, waarmee de tijdas naar wens kan worden gewijzigd. Het is heel eenvoudig om, alvorens te resynthetiseren, de stapgrootte in de tijd van de opeenvolgende frames te wijzigen, zodat zins- of woorddelen naar wens versneld of vertraagd worden weergegeven met behoud van het correcte toonhoogteverloop.
Een voor de praktijk interessante toepassing tenslotte is de mogelijkheid om spraak zeer zuinig te coderen. De modelparameters lenen zich erg goed voor drastische bezuinigingen. Het oor blijkt weinig gevoelig te zijn voor kleine 'onjuistheden' in een aantal fysische grootheden. Zonder al te grote vermindering van de spraakkwaliteit is het mogelijk om de afzonderlijke frames met in totaal slechts 30 bits te coderen. Bovendien is de stapgrootte in de tijd van ieder frame te vergroten tot 30 ms in plaats van de oorspronkelijke 10 ms. De geresynthetiseerde spraak klinkt dan wel wat minder zorgvuldig gearticuleerd, omdat hele snelle overgangen minder goed gerepresenteerd worden, maar voor veel toepassingen is dit toelaatbaar. Daarmee zijn we dan gekomen op redelijk tot goed verstaanbare spraak, die met slechts 1000 bits per seconde gecodeerd is. Vergeleken met de oorspronkelijke gedigitaliseerde versie is dat een besparing in geheugencapaciteit van een factor 120. Weliswaar worden geheugens nog steeds snel groter en goedkoper, maar in toepassingen waar een groot vocabulair vereist is, valt zo'n besparing niet te versmaden. Deze zuinige codering van spraak zal o.a. toegepast worden in een spraakchip waarin het complete synthesedeel is geïntegreerd. Dat hiermee boeiende mogelijkheden ontstaan voor toepassing van spraak als communicatiemiddel, staat vast. Hetgeen niet wegneemt dat voor een zinvolle verwezenlijking van die mogelijkheden nog zeer veel fundamenteel onderzoek vereist is, waaraan het IPO zijn bijdrage levert.

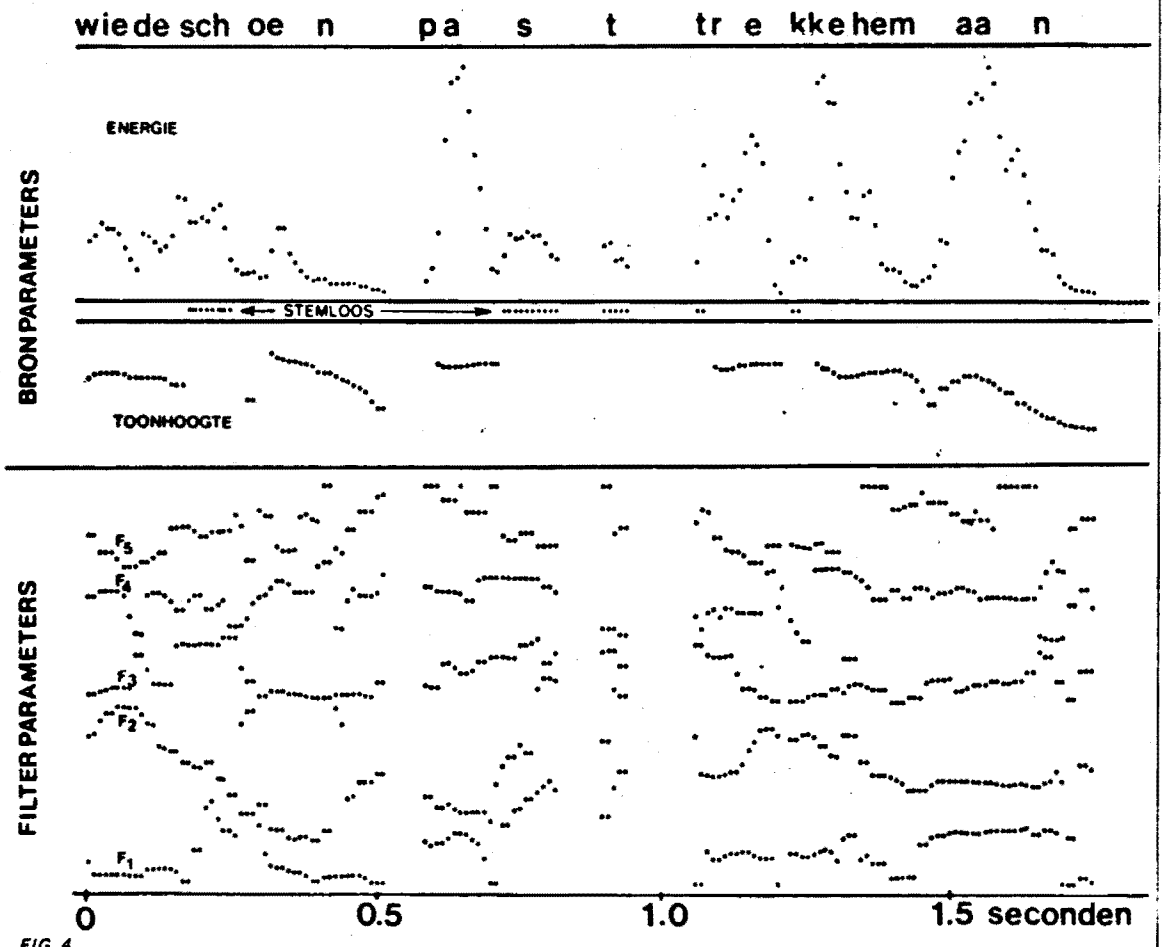

FIG. 4

Resultaat van een analyse met de rekenmachine voor het spreekwoord 'wie de schoen past trekke hem aan'. Het bovenste deel toont het verloop in de tijd van de drie model-parameters die het bronsignaal beschrijven: energie, stemhebbend/stemloos-stand en toonhoogte. Het onderste deel toont vijf van de tien geanalyseerde filterparameters uit het model, die de resonanties van de mond-keelholte beschrijven. In verticale richting maken de stippen deel uit van één frame dat correspondeert met een stapje in de tijd van 10 msec.

Uitslag verkiezingen

Op vrijdag 11 april heeft het Centraal Stembureau de uitslag bekendgemaakt van de vorige week donderdag 3 april gehouden verkiezingen van de studentleden van de hogeschoolraad en van de (onder)afdelingsraden.
De namen van degenen, die gekozen zijn volgen hieronder.

A. Hogeschoolraad
De volgende kandidaten zijn gekozen:
F.P. van Velden
W.H.J. Merkelbach
E.C. Horikx
H.J. Buitendijk
K.E.T. Reynen
J.M.G. Kleintjens
H.E. Wagter
E.G.A. Schelfaut
P.J.M. Thomassen

B. (Onder)afdelingsraden
De volgende kandidaten zijn gekozen:

Wiskunde
J.C.J.J. van der Leur
F.J.J.M. Merkx
J.H.Th.M. van Mil

Bedrijfskunde
G.M.J.M. Coppens
J.J.M. Leunissen
R.A.J.H. Thelosen

Technische Natuurkunde
J.A.M.G. Keulen
G.J.A. Hellings
C.A.M. de Vries

Werktuigbouwkunde
J. Vosmer
R. Bijl
R.J.A. Frielink
J.T.G. Gunsing

Elektrotechniek
M.H.M. Hutschemaekers
W.H.B. Bartelds
M.P.A.M. Bogers

Scheikundige Technologie
Th.W.J. van de Beek
J. Mallie
R.M.B. van Swieten

Bouwkunde
J.P.H.M. Pierey
W.G. Key
M.J.H. Philippens

Centraal Stembureau

Stellingen

Vriendjespolitiek
Democratisering van de krijgsmacht zou zich mede dienen uit te strekken tot inspraak in de bepaling wie als vijand aangemerkt moet worden.
(H.J.M. van den Bosch, Nijmegen)

Jong geleerd is .....
De oud-hollandse zegswijze 'met diploma's op zak vind ge altijd onderdak' gaat niet meer op, evenmin als 'die niet leert in zijn jeugd heeft een leven zonder vreugd'.
(H.P. van Geijn, Nijmegen)