Analyse, zuinige codering en resynthese van spraakgeluid

ANALYSE, ZUINIGE CODERING EN RESYNTHESE VAN SPRAAKGELUID

L.L.M. Vogten

ANALYSE, ZUINIGE CODERING EN RESYNTHESE

VAN SPRAAKGELUID

PROEFSCHRIFT

ter verkrijging van de graad van doctor
in de technische wetenschappen aan de Technische
Hogeschool Eindhoven, op gezag van de rector
magnificus, prof.dr. S.T.M. Ackermans, voor een
commissie aangewezen door het college van dekanen

in het openbaar te verdedigen op
vrijdag 18 november 1983 te 14.00 uur

door

LEONARDUS LAMBERTUS MARIA VOGTEN

geboren te Kerkrade

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. H. Bouma

Prof. dr. ir. J.P.M. Schalkwijk

Met dank aan ir. L.F. Willems

Dit onderzoek heeft plaatsgevonden in het
Instituut voor Perceptie-Onderzoek (IPO)
te Eindhoven.

0 INLEIDING

Het uitwisselen van informatie tussen mensen en apparatuur is de laatste jaren in snel tempo toegenomen, mede door de snelle verbreiding van digitale computers en microprocessoren. Voor het overgrote deel vindt deze communicatie thans plaats via beeldscherm, toetsenbord, printer e.d., waarbij de geschreven taal als informatiedrager fungeert. Gesproken taal, het meer natuurlijke communicatiemedium voor de mens, wordt nog vrijwel niet toegepast. Toch is tussen mensen onderling de gesproken overdracht van informatie vaak snel en efficiënt; spreken en verstaan vereisen meestal minder inspanning dan schrijven en lezen en visueel of fysiek kontakt is niet noodzakelijk. Zeker in situaties waarin ogen en handen al belast zijn, of waar visuele overdracht van informatie niet beschikbaar is, zoals bij sommige gehandicapten, zou spraak nuttige diensten kunnen bewijzen bij de communicatie met apparatuur.

Deze toepassing van spraak betekent allereerst dat apparaten spraak moeten kunnen produceren ('spreken'). Tekst die nu nog op een beeldscherm verschijnt zal omgezet moeten worden in geluiden die door de mens als spraakuitingen zijn te verstaan. Woorden en zinnen moeten liefst klinken als vloeiende en natuurlijke spraak, zodat ze snel en moeiteloos correct herkend kunnen worden. Daarnaast zal bij tweerichtingsverkeer van informatie de apparatuur spraak ook moeten kunnen verstaan, hetgeen automatische herkenning inhoudt van menselijke spraak.

Een fundamenteel probleem bij de ontwikkeling van zowel 'sprekende' als 'luisterende' apparaten is de grote variabiliteit die het geluid van menselijke spraak vertoont. Niet alleen tussen verschillende sprekers of spreeksters, maar ook bij één en dezelfde stem is de relatie tussen de waargenomen spraakklank en het bijbehorende geluid allesbehalve eenduidig (Bouma, 1976; Nooteboom en Cohen, 1976). De wijze waarop bij het spreken klanken worden gerealiseerd is sterk afhankelijk van positie, klemtoon en ook van andere spraakklanken in hun omgeving (Nooteboom, 1972; Pols, 1977; Koopmans van Beinum, 1980). Het akoestische geluidsignaal van spraak, hier verder afgekort met spraaksignaal is, in de vorm van luchtdrukveranderingen in de tijd, de fysische drager van spraak. Het is zowel het uitgangssignaal van het spraakproduktieproces als het ingangssignaal voor het proces van herkenning waarop spraakverstaan berust. De golfvorm van het spraaksignaal, dit is de momentane waarde van de luchtdruk als funktie van de tijd, is relatief gemakkelijk fysisch te registreren, op te slaan en weer te geven in tijd- en frekwentiedomein, analoog of digitaal. Het spraaksignaal staat centraal in de experimentele fonetiek (Nooteboom

en Cohen, 1976), een onderzoekterrein dat ons kennis en inzicht kan verschaffen die onmisbaar is voor de ontwikkeling van apparatuur waarin gesproken taal als informatiedrager fungeert. Een sprekend apparaat moet immers geluidssignalen opwekken met zodanige fysische eigenschappen dat de luisteraar deze als spraakklanken waarneemt, herkent en verstaat. Zo ook zal het herkenningsproces van een 'luisterend' apparaat moeten opereren op de fysische eigenschappen van het menselijk spraaksignaal.

0.1 PROBLEEMSTELLING

We laten de problematiek van automatische spraakherkenning door 'luisterende' apparaten verder rusten en beperken ons tot spraakuitgifte. Daarbij denken we alleen aan zakelijke, informatieve of beschrijvende uitingen. Stemming, gemoedstoestand en dergelijke faktoren laten we daarmee buiten beschouwing. Met sprekende apparaten bedoelen we hier meer dan gangbare geluidsregistratie met apparatuur die gebruik maakt van magneetband, plaat of schijven als opslagmedium. Immers naar analogie van tekst op een beeldscherm moeten willekeurige, dat betekent een praktisch onbegrensd aantal, uitingen kunnen worden samengesteld uit een vaste, begrensde en liefst niet te grote verzameling eenheden. Bij geschreven taal zijn dat de letters en leestekens; bij gesproken taal zouden dat in principe de kleinste spraaksegmenten kunnen zijn die verschillen in betekenis dragen: de fonemen. Een groot probleem hierbij is echter de al genoemde variabiliteit van het spraaksignaal van (in dit geval) de fonemen. De fysische eigenschappen van deze eenheden zijn in gesproken taal sterk afhankelijk van de context waarin ze worden uitgesproken en variëren voortdurend binnen één spraakklank. Fonemen hebben dus geen vaste fysische vorm binnen het woord waarin ze voor komen. Die omgevingsafhankelijkheid in rekening brengen door grotere eenheden te gebruiken, zoals foneemovergangen, lettergrepen of woorden, levert geen oplossing. Zelfs op het nivo van de langste eenheden, woorden, geldt dat het reproduceren en zonder meer aan elkaar rijgen van de spraaksignalen meestal geen vloeiende spraak oplevert.

Waargenomen spraakklanken zijn dus fysisch niet vormvast en voor het spraakonderzoek is het daarom van groot belang dat de afzonderlijke fysische eigenschappen beheerst en gevarieerd kunnen worden ('t Hart et al, 1981/1982). Pas dan kan worden nagegaan welke van die eigenschappen relevant zijn en als kenmerk fungeren voor de menselijke spraakperceptie, het proces van horen, herkennen en verstaan van spraakklanken. Voorbeelden van fysische eigenschappen waarmee ieder geluid en dus ook spraakgeluid kan worden beschreven zijn: tijdsduur, intensiteit, amplitude- en fasespectrum en periodiciteit. Ieder van

deze eigenschappen kan op zijn beurt door een of meer parameters worden gespecificeerd.

In het algemeen is de fysische samenstelling van het natuurlijke spraaksignaal vrij ingewikkeld als funktie van de tijd. Er zijn echter eigenschappen waarvan lang niet alle details in het tijd- of frekwentiedomein noodzakelijk zijn voor correcte herkenning van spraakklanken of om die spraak als vloeiend en natuurlijk waar te nemen. In het spraakonderzoek in het algemeen, en bij de ontwikkeling van sprekende apparatuur in het bijzonder, streeft men er dan ook naar om die ingewikkelde samenstelling zó ver te vereenvoudigen dat alleen datgene overblijft wat nodig is om aan de gestelde eisen voor die uitgifteapparatuur te voldoen. Naarmate die eisen hóger liggen en b.v. niet alleen goed verstaanbare maar ook natuurlijk klinkende spraak is gewenst, zullen er in principe ook meer details in het verloop van de fysische eigenschappen bewaard moeten blijven. Maar welke dat dan zijn is nog deels onbekend en zal door experimenteel onderzoek moeten worden vastgesteld. Ook hiervoor is nodig dat afzonderlijke parameters onafhankelijk kunnen worden gevarieerd zodat het effekt daarvan op de spraakperceptie kan worden vastgesteld.

Bij het onderzoek aan spraak wordt veel gebruik gemaakt van kunstmatig opgewekte, 'spraakachtige' signalen: synthetische spraak. Dat gebeurt dan met behulp van een model dat een, al dan niet vereenvoudigde, beschrijving geeft van de fysica van de articulatie, het laatste gedeelte van het proces van de natuurlijke spraakproduktie via de menselijke stem. Essentieel voor dit soort synthetische spraak is dat aan de produktie ervan een model ten grondslag ligt, waarvan de parameters, afhankelijk van het model, meer of minder direkt fysische eigenschappen van het natuurlijke spraaksignaal representeren. Spraaksignalen waarvan b.v. de golfvorm alleen maar (zuinig) gecodeerd is (pulscode- of deltamodulatie), vallen dus niet onder het begrip synthetische spraak. Wel valt eronder het uitgangssignaal van een bronfiltermodel (waarover meer in het volgende hoofdstuk), waarin het filter een fysische representatie is van de akoestiek van de mond-keelholte.

Onderzoek naar een zuinige beschrijving van het spraaksignaal via het gebruik van synthetische spraak komt dan in feite neer op het specificeren van een (hanteerbaar) model en het vinden van de parameterwaarden daarvan als funktie van de tijd. Synthetische spraak zal dan beter met natuurlijke spraak overeenkomen naarmate het gehanteerde produktiemodel het articulatieproces beter weerspiegelt en de modelparameters beter de fysische parameters van natuurlijk spraakgeluid representeren. De vraag is dan hoe we de parameters van het model kunnen bepalen. Daarbij kunnen we twee werkwijzen onderscheiden.

De eerste mogelijkheid is om uit te gaan van de allereenvoudigste en

kortste spraakklanken, fonemen, en daarvoor de modelparameters zo te bepalen dat de synthese daarvan goed klinkt. Om daarmee vervolgens grotere spraakuitingen samen te stellen moet dan een receptuur van context-afhankelijke regels worden opgesteld om de eerdergenoemde variabiliteit in het spraaksignaal in rekening te brengen. Deze methode heet dan ook 'synthese door regels'. Naarmate de spraakuiting langer wordt nemen omvang en ingewikkeldheid van de receptuur snel toe; zij is bovendien sterk taalafhankelijk. Voor het amerikaans en brits engels bestaan er enkele spraakuitgiftesystemen die van deze methode gebruik maken (Klatt, 1976). De geproduceerde spraak klinkt redelijk en is doorgaans goed verstaanbaar maar verschilt perceptief nog vrij sterk van natuurlijke spraak. Ook voor het nederlands is zo'n systeem van regels opgesteld (Slis en Muller, 1971, Slis et al, 1977).

De tweede mogelijkheid is om te beginnen bij komplete, eventueel lange, natuurlijke spraaksignalen en daaruit de parameters van het model te bepalen via een geschikte fysische analyse. Spraak die via deze methode van 'analyse-resynthese' wordt verkregen klinkt thans veel natuurlijker dan via synthese door regels, omdat we uitgaan van natuurlijke spraak. De volgende stap is dan om door vereenvoudiging van de verkregen analyseresultaten een zuinige beschrijving te zoeken en op deze manier voorschriften te vinden waaraan de fysische eigenschappen moeten voldoen om goed klinkende spraak te verkrijgen. Over deze laatstgenoemde methode van analyse-resynthese handelt het onderzoek dat hier zal worden beschreven.

Doel van dit onderzoek vormt de ontwikkeling en realisatie van een flexibel toepasbaar instrument voor het spraakonderzoek, waarmee onderzocht kan worden welke rol de fysische parameters van het spraakgeluid spelen in de spraakperceptie. Aan dit systeem stellen we in principe de volgende eisen:

1. analyse en resynthese vinden plaats in termen van een beperkt aantal fysische parameters die perceptief relevant zijn.
2. de synthetische spraak die met het systeem wordt geproduceerd is perceptief niet te onderscheiden van de oorspronkelijke spraak.
3. de parameters zijn snel en automatisch rechtstreeks uit het spraaksignaal zelf te bepalen.
4. de parameters zijn geschikt om zuinig te coderen voor toepassing in praktisch realiseerbare systemen voor spraakuitgifte door apparaten.

In dit boekje zullen we principe, uitvoering, resultaten en toepassingen bespreken van het door ons ontwikkeld systeem voor analyse en resynthese van spraak.

De fundamenten van dit systeem zijn indertijd gelegd door Willems (1976), in de vorm van programma's waarmee spraakinname, -analyse en

uitgifte kon worden gerealiseerd met de toenmalige IPO P9202 computer. Hiermee werden in beginsel de parameters berekend voor het besturen van een (hardware) spraaksynthese-apparaat (Vogten en Willems, 1977). Sinds begin 1978 is deze hardwaresynthese vervangen door software, waarmee aanzienlijk betere kwaliteit van de geproduceerde spraak kon worden bereikt. Daarna zijn analyse en synthese voortdurend verbeterd, uitgebreid en aangepast aan eisen voor gebruik in het spraakonderzoek. Mogelijkheden voor beperkt interaktief gebruik werden voor het eerst gerealiseerd op de IPO P857 minicomputer maar sinds de komst eind 1981 van de snellere en krachtiger VAX 11/780 staat een uitgebreid pakket programmatuur ter beschikking voor interaktief spraakonderzoek. Het systeem bestaat in hoofdzaak uit Fortran programma's, waarvan de belangrijkste zijn opgesomd in Vogten (1983).

In essentie is dit analyse-resynthesesysteem gebaseerd op een, in de experimentele fonetiek algemeen aanvaard, model voor de spraakproduktie, het bron-filtermodel van Fant (1960). Dit model beoogt een sterk vereenvoudigde weergave te zijn van de fysica van de eindfase van de menselijke spraakproduktie, waarin de eigenlijke spraakklanken mechanisch/akoestisch worden opgewekt.

We zullen daarom in hoofdstuk 1 eerst deze fysica van de menselijke spraakproduktie in het kort bespreken, daarna het daarop gebaseerde principe van het bron-filtermodel uiteenzetten en dan het door ons toegepaste model specificeren. In vergelijking met de fysica van de menselijke spraakproduktie is dat model een sterke vereenvoudiging en het vertoont door zijn beperkingen ook duidelijke verschillen daarmee. Toch zullen we dit simpele model toepassen omdat door die beperkingen en eenvoud de parameters van dit bron-filtermodel automatisch, snel en rechtstreeks uit de golfvorm van het spraaksignaal zelf kunnen worden berekend.

Hoe we dat doen wordt in hoofdstuk 2 uiteengezet, waarin we tevens laten zien hoe de modelparameters na berekening vertaald kunnen worden in parameters die nauw aansluiten bij relevante grootheden in de spraakperceptie.

Met de resultaten die uit deze analyse zijn verkregen kan het spraaksignaal vervolgens weer gereconstrueerd worden, geheel conform het bron-filtermodel. Het principe en de praktische uitvoering van deze resynthese worden behandeld in hoofdstuk 3. Het aldus geresynthetiseerde signaal vertoont veel overeenkomsten met het oorspronkelijke spraaksignaal maar ook verschillen, mede vanwege de genoemde modelbeperkingen.

De gevolgen van die beperkingen komen in hoofdstuk 4 aan de orde, waarin, bij wijze van fysische evaluatie van het systeem, wordt besproken in hoeverre de modelbeperkingen alsmede de keuze van de

vaste modelparameters van invloed zijn op de fysische eigenschappen van de geresynthetiseerde spraak.

De perceptieve gevolgen van de modelbeperkingen en van de keuze van de vaste modelparameters komen vervolgens in hoofdstuk 5 aan de orde. Hierin gaan we na in hoeverre individuele spraakklanken, met name medeklinkers, door het analyse-resynthesesysteem worden aangetast en door de luisteraar niet meer correct worden herkend. Bij deze perceptieve evaluatie beperken we oms om praktische redenen in hoofdzaak tot de herkenning van medeklinkers in losse woordjes. We laten daarbij allerlei andere aspecten van de synthetische spraak zoals herkenbaarheid van de spreker, ritmiek, mate van natuurlijkheid e.d. buiten beschouwing. Ritmiek en sprekerherkenbaarheid worden door het systeem overigens niet of nauwelijks aangetast en voor het bepalen van de natuurlijkheid van geresynthetiseerde spraak zijn nog geen geschikte methodes beschikbaar.

Hoofdstuk 6 geeft vervolgens een kort overzicht van een aantal toepassingsmogelijkheden die het systeem biedt voor experimenteel onderzoek aan spraak, zowel theoretisch als praktijkgericht, met name voor de ontwikkeling van apparatuur voor zuinige codering en uitgifte van spraak. Belangrijk kenmerk van het systeem is daarbij dat manipulaties, vereenvoudigingen en bezuinigingen in het parameterbestand snel, flexibel en interaktief zijn uit te voeren, waarbij de perceptieve gevolgen van deze ingrepen direkt zijn te beluisteren via de geresynthetiseerde spraak.

Hoewel ons analyse-resynthesesysteem in ruime mate voldoet aan de hierboven gestelde eisen zijn er toch nog duidelijk zwakke plekken aan te wijzen. Welke dat zijn en hoe de door het systeem gegenereerde spraak verder verbeterd zou kunnen worden, komt aan de orde in hoofdstuk 7, waarin we enkele mogelijkheden aangeven voor toekomstig onderzoek.

1 EEN MODEL VOOR MENSELIJKE SPRAAK-PRODUKTIE

Mensen spreken doorgaans met de intentie om via spraakuitingen een gedachte, mededeling of bedoeling over te brengen aan anderen. In het brein van de spreker wordt die uiting in de vorm gegoten van woorden en zinnen uit een bepaalde (natuurlijke) taal. Normaal gesproken vinden dan, onder controle van datzelfde brein, talloze gecoördineerde spiercontracties plaats, die middenrif, ribben, strottehoofdbeentjes, kaken, verhemelte, tong en lippen zodanig doen bewegen dat de bedoelde klanken ontstaan (Nooteboom en Cohen, 1976).

Alleen op het laatste stadium van het zeer ingewikkelde proces van de spraakproduktie zullen we in het bestek van dit hoofdstuk iets dieper ingaan. We beperken ons daarbij tot het mechanisch/akoestische gedeelte van het spraakproduktieproces; dat gedeelte waarin de klanken worden opgewekt door bewegingen van de spraakorganen. Alles wat aan dit fysische stadium voorafgaat en nodig is om die gecoördineerde spierbewegingen uit te voeren laten we hier verder buiten beschouwing.

In de volgende paragraaf zullen we eerst de fysica van de spraakproduktie in het kort toelichten, ter inleiding op een daarna te bespreken eenvoudig model voor de spraakproduktie. Dit model beoogt de belangrijkste elementen van de fysica der spraakproduktie weer te geven en staat dan ook centraal in ons analyse-resynthesesysteem. We besluiten het hoofdstuk met een specificatie van dit door ons toegepaste model en geven de belangrijkste verschillen die het vertoont met de fysica van de menselijke spraakproduktie.

1.1. FYSICA VAN DE SPRAAKPRODUKTIE

Fysisch gezien zijn spraakklanken hoorbare luchtdrukveranderingen die teweeg worden gebracht door het mechanisme van de menselijke stem. Bij klanken van normale (nederlandse) spraak worden deze veranderingen opgewekt doordat uitgeademde lucht ergens in de mond-keelholte een vernauwing passeert. Wanneer die vernauwing optreedt bij de stemspleet, de ruimte tussen de stembanden in het strottehoofd, kunnen de stembanden periodiek open en dicht klappen, waarbij relatief snel opeenvolgende geluidsplofjes ontstaan. Dan ontstaan stemhebbende klanken, zoals klinkers en tweeklanken. Door spiertjes van en bij de stembanden in het strottehoofd te spannen of te ontspannen kan de trillingsfrekwentie van de stembanden binnen zekere grenzen verhoogd of verlaagd worden. De luchtdrukveranderingen die door de stembandtrillingen worden veroorzaakt hebben bij benadering een driehoekig verloop, waarbij, afhankelijk van de duur van een periode, ruwweg de helft van de periode de stemspleet geheel gesloten is. De opgewekte geluidsenergie is dus telkens geconcentreerd in een vrij korte tijdsduur, zodat het energiespectrum zich over een groot frekwentiegebied uitstrekt. De stembandtrillingen bevatten dus veel hogere harmonischen, waarvan de amplitude in eerste benadering afneemt met het kwadraat van de frekwentie. De omhullende van dit energiespectrum heeft dus een helling van ongeveer -12 dB/octaaf.

Door de akoestische eigenschappen (resonanties, absorpties en reflekties) van de keel-, mond- en neusholtes en lippen, hier verder het 'mondkanaal' genoemd, wordt het spectrum van de stembandtrillingen gefilterd, 'gekleurd' tot een spectrum met een veel grilliger gevormde omhullende. De stand van vooral tong en onderkaak bepalen daarbij, door plaatselijke resonanties, in welke frekwentiegebieden de trillingen verzwakt of relatief versterkt worden. In het energiespectrum van de aldus gevormde klanken zijn dan frekwentiegebieden te onderscheiden waar de energie relatief hoog is. Die gebieden zijn karakteristiek, vooral voor afzonderlijke klinkers en tweeklanken en wel zó karakteristiek dat ze al sedert bijna een eeuw de naam 'formanten' dragen. Voorbeelden van de korte-termijn energiespectra van (los ingesproken) klinkers ee en e zijn in fig. 1.1 weergegeven, tezamen met de bijbehorende golfvorm, dat is het verloop van de drukveranderingen in de tijd.

Formanten worden dus gevormd door het mondkanaal, waarvan de vorm op zijn beurt de spraakklanken formeert. Zij werden al in 1889 door Hermann "maatgevende bestanddelen van klinkers" genoemd en Stumpf vond dat ze de "toonkwaliteit bepalen en in hoge mate bijdragen tot het karakter van een klinker" (ontleend aan Chiba en Kajiyama, 1958). Later is vaak bij de omschrijving van het begrip formant een koppeling

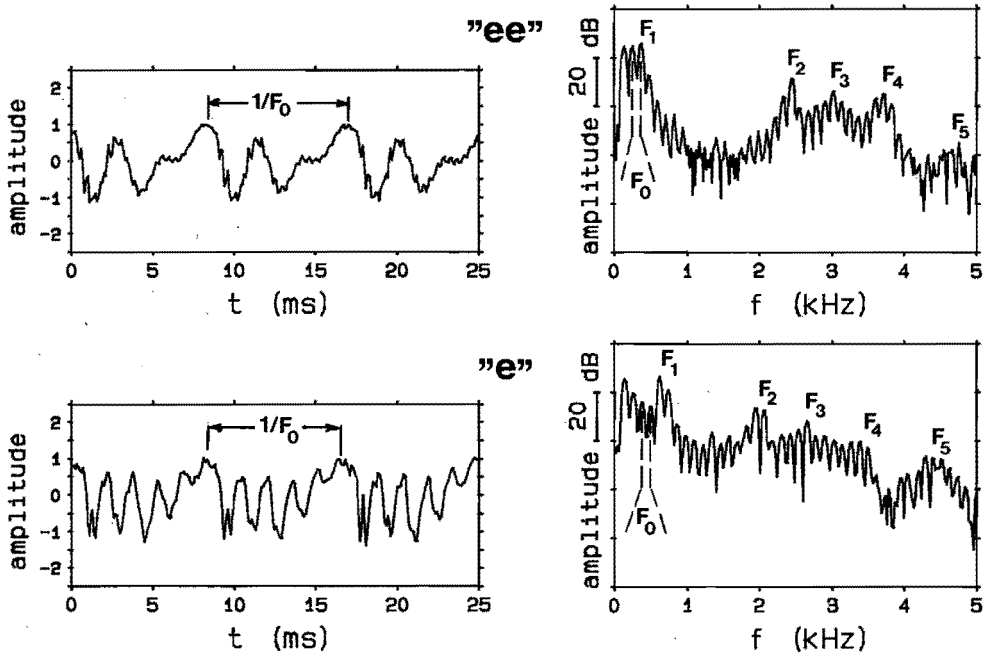

Fig.1.1. Voorbeelden van golfvorm (links) en energiespectrum (rechts) van los ingesproken klinkers. Boven: lange ee als in "keet", onder: korte e als in "pet". De herhalingsfrekwentie van de stembandtrillingen (hier ongeveer 110 Hz) vinden we terug in de periode $1/F_0$ van de golfvorm en in de fijnstruktuur van het spectrum. In de spectra zijn 5 formanten te herkennen, aangegeven met F_1 t/m F_5.

gelegd met de omhullende in het energiespectrum, die de toppen van de afzonderlijke harmonischen met elkaar verbindt. Formanten manifesteren zich vaak bij frekwenties waar die omhullende maximaal is (Fant, 1968). In het voor de spraakperceptie belangrijke frekwentiegebied tussen 0 en 5 kHz worden doorgaans echter niet meer dan 5 formanten onderscheiden. Dus lang niet alle spectrale toppen in het natuurlijke spraaksignaal worden formanten genoemd; formanten zijn niet eenduidig uit het spraaksignaal zelf te bepalen.

Voor gewone klinkers en tweeklanken hebben we gezien hoe resonanties in het mondkanaal de energie in bepaalde delen van het spectrum verhogen. Bij nasale medeklinkers, zoals bv. m, n en ng, is de mondopening afgesloten en zijn er naast maxima in het energiespectrum ook gebieden te vinden waar de energie juist laag is ('nulpunten'). Daarvan zien we voorbeelden in fig. 1.2. Meestal uit zo'n nulpunt zich alleen in een steilere helling die zich over grotere energienivo's uitstrekt dan bij

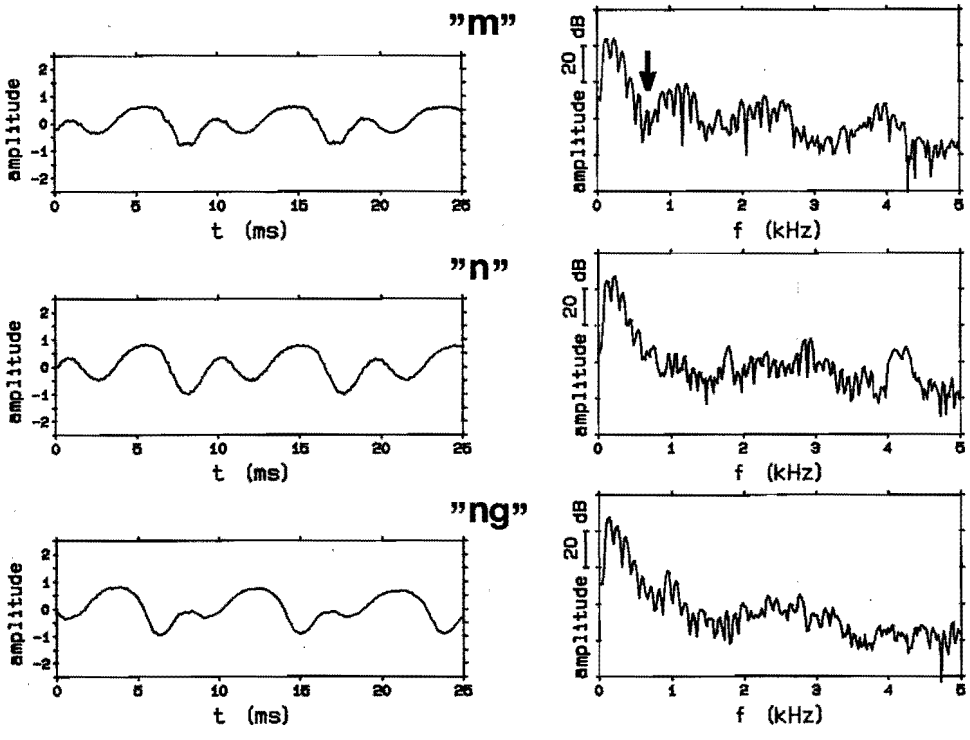

Fig.1.2. Voorbeelden van golfvorm (links) en energiespectrum (rechts) van nasalen. Van boven naar beneden: m uit "naam", n uit "maan" en ng uit "jong". Bij de m is een nulpunt in het spectrum te herkennen bij ongeveer 600 Hz, aangegeven met de pijl.

gewone klinkers het geval is. Ook nulpunten zijn niet eenduidig te bepalen uit het spraaksignaal.

Hoewel de stembanden een belangrijke rol spelen bij het produceren van spraakklanken vormen zij niet de enige geluidsbron. Ook op andere plaatsen in het mondkanaal kunnen vernauwingen leiden tot het ontstaan van hoorbare geluidstrillingen, die dan echter doorgaans niet periodiek zijn. Deze wrijfklanken ontstaan door snelle wervelingen (turbulenties) in de uitgeademde luchtstroom en hebben van oorsprong eveneens een breed energiespectrum, dat dan ook weer door akoestische resonanties in het mondkanaal meer of minder gefilterd kan worden, afhankelijk van de plaats waar zo'n vernauwing optreedt. Is die plaats achter in de mond, zoals bij de ch van "schoof", dan spelen resonanties van de mondholte in belangrijke mate mee. Wanneer de ruis wordt gevormd tussen boventanden en onderlip, zoals bij de f, dan speelt de akoestische filtering van het mondkanaal geen of een veel kleinere

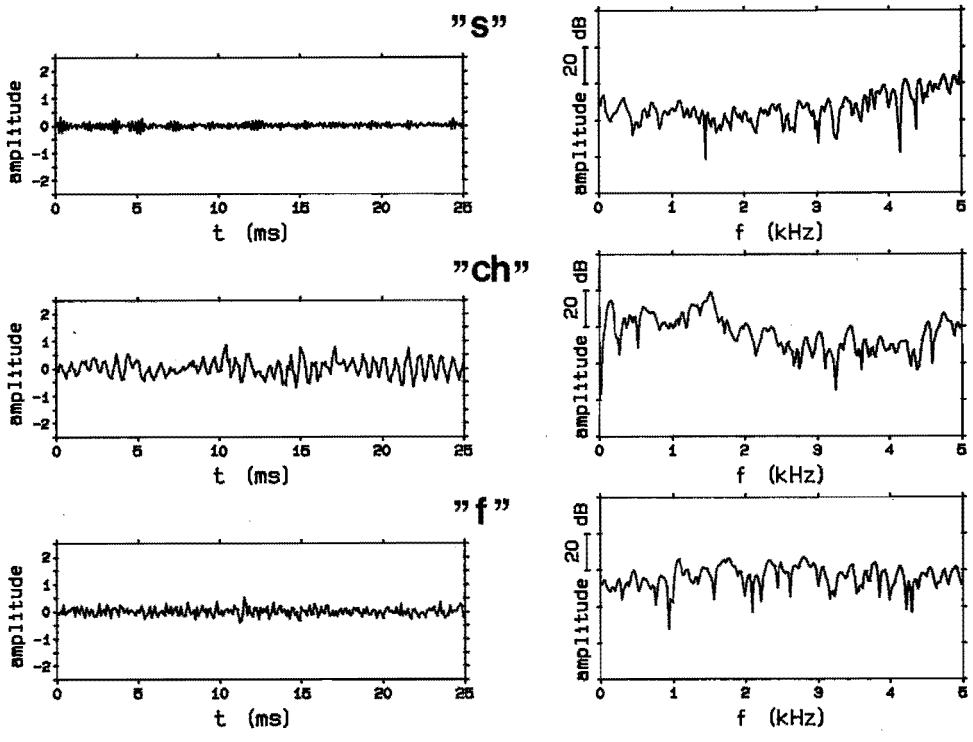

Fig.1.3. Voorbeelden van golfvorm (links) en energiespectrum (rechts) van los ingesproken stemloze wrijfklanken. Van boven naar beneden: s, ch en f.

---

rol. Naast deze stemloze wrijfklanken komen ook stemhebbende wrijfklanken voor, zoals z en v van 'zeven', waarin kombinaties van ruisige en periodieke geluidsbronnen optreden. Van beide typen wrijfklanken zien we voorbeelden in fig. 1.3 en 1.4.

Dan is er nog een derde type brongeluid dat onstaat door het abrupt opheffen van een totale afsluiting van het mondkanaal waarbij eveneens turbulenties optreden. Totale afwezigheid van brongeluid tijdens de opbouwfase van de druk en een daarop volgend ruisplofje treedt op bij de stemloze plofklanken p, t en k. Iets soortgelijks gebeurt bij de stemhebbende plofklanken b, d en zachte k van 'zakdoek', waarbij dan echter tijdens de opbouwfase van de druk de stembanden wel hoorbaar trillen (Nooteboom en Cohen, 1976).

Bij het produceren van al deze klanken heeft de spreker ook nog de mogelijkheid om via de ademhalingsspieren de luchthoeveelheid die per seconde door de keelholte stroomt te beïnvloeden. Hiermee kan zowel de

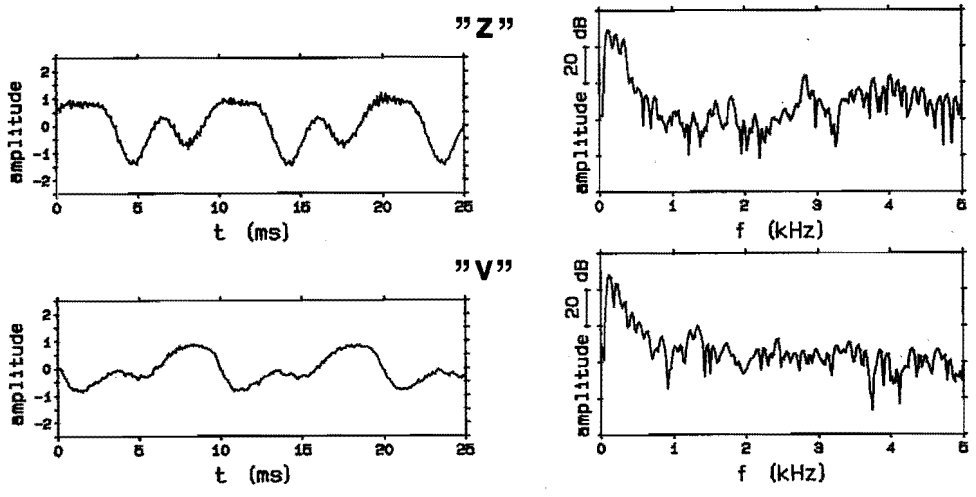

Fig.1.4. Voorbeelden van golfvorm (links) en energiespectrum (rechts) van los ingesproken stemhebbende wrijfklanken. Boven: z, onder: v. Vooral bij de z varieert de ruis binnen de grondtoonperiode sterk in amplitude. In het spectrum zijn grote gebieden zonder duidelijke periodieke struktuur.

amplitude van de stembandtrillingen alsook de energie van de ruisgeluiden en daarmee dus de energie van het spraaksignaal worden verhoogd of verlaagd.

Wanneer de trillende lucht uiteindelijk via de mond en/of de neus naar buiten stroomt treedt in dit laatste stadium nog het stralingseffekt op van de uitstroomopening. De lage harmonischen worden daarbij relatief meer verzwakt dan de hoge en het energiespectrum ondergaat een extra verandering overeenkomend met een helling van ongeveer +6 dB/oct.

Datgene wat wij als lopende spraak waarnemen is in feite een aaneenschakeling van bovengenoemde klanken, soms in onderlinge kombinatie en meestal zonder duidelijke, fysisch waarneembare, grenzen tussen de opeenvolgende klanken. We zien daarvan een voorbeeld in fig. 1.5, waarin de golfvorm van een stuk lopende spraak is weergegeven. Dit voorbeeld illustreert tevens iets van de grote variabiliteit van het spraaksignaal, die hier o.m. tot uiting komt in het verschil (in vooral de amplitude) tussen de ee van "weet" en die van "heeft". Ook de d van "wie de" is totaal verschillend van de d in "gevonde(n)"; bij de laatste is het plofachtige karakter nauwelijks of niet in de golfvorm terug te vinden.

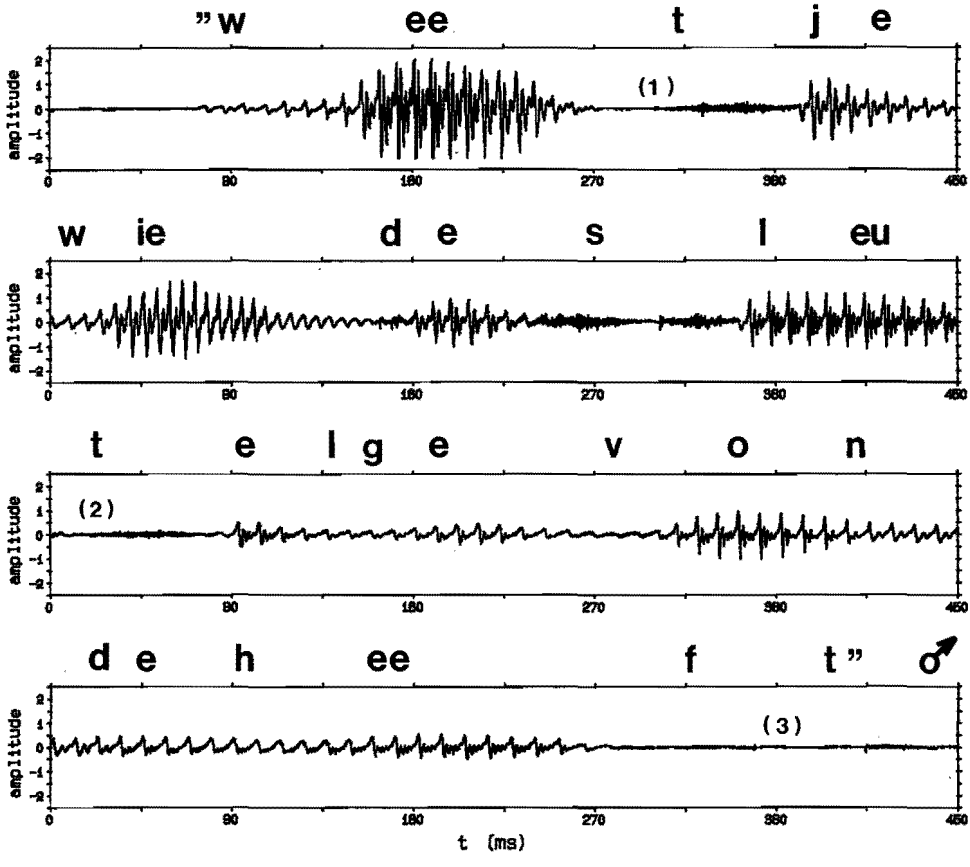

Fig.1.5. Golfvorm van de zin "weet je wié de sleutel gevonden heeft", uitgesproken door een mannenstem. Periodieke en ruisige signalen wisselen elkaar af en er zijn geen duidelijke, fysisch waarneembare grenzen aan te wijzen tussen de woorden onderling. In de pauzes vóór de plofklank t van "weet", "sleutel" en "heeft", aangegeven door resp (1), (2) en (3), is het signaal niet helemaal nul vanwege o.a. bandruis en kwantiseringsruis.

Bij het spreken veranderen de akoestische eigenschappen van zowel geluidsbron(nen) als de 'aangeslagen' ruimtes in het mondkanaal, het akoestisch filter, voortdurend in de tijd. De fysische eigenschappen van het signaal dat wij als spraak kunnen waarnemen veranderen daarmee eveneens in de tijd. Daarbij zijn enkele eigenschappen te onderscheiden die een vrij direkte binding hebben met concrete perceptieve grootheden.

1. De amplitude c.q. de energie van het periodieke c.q. ruisige signaal, die de intensiteit bepaalt en daarmee vooral (maar niet al-

leen) de waargenomen luidheid van de klank.
2. De periodiciteit, die bepaalt of de klank als stemhebbend, stemloos of als een kombinatie van beide wordt waargenomen.
3. De grondtoon van een periodiek signaal, die nauw samenhangt met waargenomen toonhoogte. Het verloop van de toonhoogte als funktie van de tijd wordt aangeduid met intonatie.
4. De omhullende van het korte-termijn energiespectrum van het spraaksignaal. Daarin is vaak een bepaalde formantstruktuur te onderscheiden en de vorm van die omhullende hangt samen met de waargenomen timbre of 'klankkleur', een kenmerk waarin o.a. de verschillende klinkers zich nog van elkaar onderscheiden wanneer hun luidheid en toonhoogte niet van elkaar verschillen. Veranderingen in het korte-termijn energiespectrum, en dus van de formanten, als funktie van de tijd, spelen onder meer een belangrijke rol bij de waarneming van tweeklanken en van medeklinkers. De vorm van het energiespectrum is eveneens (mede)bepalend voor de waargenomen luidheid.

Hiermee hebben we de belangrijkste fysische aspecten van de menselijke spraakproduktie aangegeven, tezamen met hun perceptieve tegenhangers en gaan we over tot de bespreking van een fysisch model hiervoor.

1.2. BRON-FILTERMODEL VOOR SPRAAKPRODUKTIE

In de voorgaande fysische beschrijving van de menselijke spraakproduktie zijn termen als 'brongeluid' en 'akoestisch filter' gebruikt. Daarmee is de essentie van het bron-filtermodel al aangeduid. Dit model beoogt een (vereenvoudigde) beschrijving te geven van de fysica van de menselijke spraakproduktie. Dat wil zeggen dat met dit model spraakachtige signalen kunnen worden geproduceerd waarin de belangrijkste fysische eigenschappen door de modelparameters worden gerepresenteerd. Het model beschrijft deze eigenschappen als funktie van de tijd door middel van een variabel bronsignaal dat als ingangssignaal dient voor een eveneens variabel lineair filter. Een belangrijke eigenschap is dat bron en filter wederzijds onafhankelijk zijn en elkaar niet belasten. In het tijddomein is het uitgangssignaal van het model dan de convolutie van bronsignaal en impulsresponsie van het filter. In het frekwentiedomein is het uitgangsspectrum het produkt van bronspectrum en overdrachtsfunktie van het filter. Deze overdrachtsfunktie bepaalt de omhullende en het bronsignaal de fijnstruktuur van het energiespectrum van het uitgangssignaal.

Het bronsignaal beschrijft het akoestische signaal van de periodiek trillende stembanden of van de luchtwervelingen in het mondkanaal als

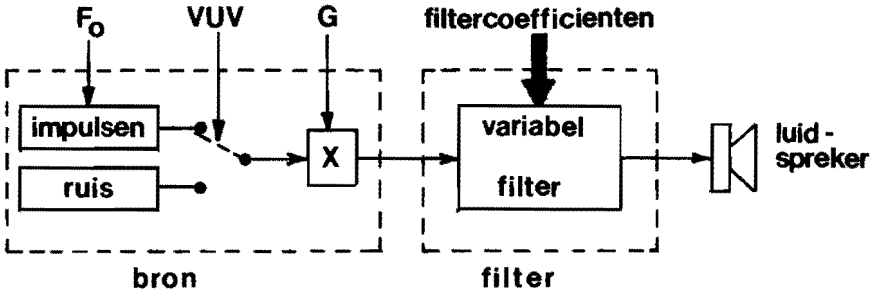

Fig.1.6. Bron-filtermodel voor de produktie van spraakgeluid. Het bronsignaal wordt bepaald door drie parameters: de herhalingsfrekwentie F_0 van de impulsen, de stem/stemloosparameter VUV en de versterkingsfaktor G.

---

funktie van de tijd. Voor stemhebbende stukken spraak is dit bronsignaal een periodieke impuls met herhalingsfrekwentie F_0, die de stembandklapjes representeert. Voor stemloze spraaksegmenten bestaat het bronsignaal uit ruis. Welke van deze twee bronnen als ingangssignaal voor het filter dient, bepaalt een binaire stem/stemloos-parameter VUV (fig. 1.6). Als derde bronparameter is in het model een variabele versterkingsfaktor G opgenomen, waarmee de amplitude van het bronsignaal als funktie van de tijd wordt gerepresenteerd.

Het filter beschrijft het akoestisch filter dat de overdracht van bronsignaal naar 'spraaksignaal' aan de uitgang van het model weergeeft. In de overdrachtsfunktie hiervan zijn een drietal komponenten ondergebracht:

1. Een overdrachtsfunktie met een (voor stemhebbende klanken) vaste, tijdsonafhankelijke helling van -12 dB/oct. We hebben in par. 1.1 gezien dat de drukveranderingen t.g.v. de stembandtrillingen in eerste benadering driehoekig van vorm zijn en dat dit energiespectrum voor toenemende frekwentie afneemt met 12 dB/oct. Deze komponent representeert het spectrale verschil tussen een impuls en de driehoekige puls van de (benaderde) stembandtrillingen.
2. De belangrijkste komponent: een variabele overdrachtsfunktie die de akoestische eigenschappen van het mondkanaal als funktie van de tijd beschrijft. Hoe we deze overdrachtsfunktie specificeren komt in volgende paragrafen aan de orde.
3. Een overdrachtsfunktie met een vaste, tijdsonafhankelijke helling van +6 dB/oct die het effekt van de uitstraling aan de mondopening representeert.

Vaak worden in de literatuur (Fant, 1968; Flanagan, 1972) (1) en (3) als afzonderlijke filters beschouwd, omdat (1) direct gekoppeld is aan de periodiekbron. Het zijn echter alle drie lineaire filters en de

resulterende overdrachtsfunktie wordt gegeven door het produkt van de afzonderlijke bijdragen. Zij mogen dus als een geheel worden opgevat en we zullen in het model ook verder geen onderscheid meer maken tussen deze afzonderlijke komponenten. Daarmee wordt dan tevens ondervangen dat komponent (1) alleen bij stemhebbende klanken aanwezig is en dus striktgenomen ook tijdafhankelijk is. In de overdrachtsfunktie van het filter zijn dus (1) het effekt van de afnemende hogere harmonischen van de stembandtrillingen, (2) de variabele akoestische eigenschappen van de keel-, mond- en neusholtes en (3) het effekt van de straling aan de mondopening opgenomen. Dit betekent dat in het model het ingangssignaal van het filter een vlak, ('wit') energiespectrum heeft, zowel voor het beschrijven van stemhebbende als stemloze klanken. De periodiekbron van het model levert dus voor stemhebbende klanken een reeks eenheidsimpulsen, met herhalingsfrekwentie F_O en de ruisbron levert voor stemloze klanken ongecorreleerde stationaire witte ruis.

Het bronsignaal in het model wordt dus in de tijd gespecificeerd door drie parameters: ruisbron of periodiekbron (VUV), herhalingsfrekwentie F_O van de periodieke impuls en versterkingsfaktor G voor de amplitude van het bronsignaal.

Hoeveel parameters er nodig zijn voor de specificatie van het filter hangt voornamelijk af van de eisen die aan het systeem en dus aan het daarin toegepaste model worden gesteld. Het model moet in principe synthetische spraak kunnen produceren die niet te onderscheiden is van de oorspronkelijke spraak. Het aantal parameters waarmee de overdrachtsfunktie van het filter wordt gespecificeerd zal dan ook ten eerste afhangen van het frekwentiegebied waarover die oorspronkelijke spraak zich in feite uitstrekt. In principe kan dat het gehele hoorbare gebied zijn tussen 0 en 20 kHz, maar in het verdere verloop van dit boekje beperken we dit gebied tot 5 kHz, omdat componenten boven 5 kHz weinig bijdragen tot de verstaanbaarheid en de kwaliteit van het spraaksignaal. Met 'natuurlijke' of 'oorspronkelijke spraak' bedoelen we hier dan ook verder steeds tot 5 kHz bandbegrensde spraak, tenzij expliciet anders is vermeld.

Verder is het aantal filterparameters afhankelijk van de gewenste mate van perceptieve overeenkomst tussen oorspronkelijke spraak en resynthese. Daarnaast willen we voor toepassing in apparatuur voor spraakuitgifte graag een zo zuinig mogelijke beschrijving van spraaksignalen geven, waarbij de eis van perceptieve overeenkomst met het oorspronkelijk signaal misschien wat minder zwaar weegt.

Bij overigens gelijke eisen zijn er nog zuinige en minder zuinige realisaties mogelijk. Zo wordt in de klassieke kanaalvocoder het frekwentiegebied van het spraaksignaal opgesplitst in een stuk of 20 vas-

te banden of kanalen, waarvoor per kanaal de amplitude wordt gespecificeerd, evt. na meting uit het spraaksignaal zelf via bandfilters. Daarmee wordt dan de overdrachtsfunktie van het filter voor discrete frekwenties beschreven. Zuiniger is een realisatie waarin het filter met deelfilters wordt beschreven, die continu afstembaar zijn langs de frekwentie-as. Een cascade van dergelijke filters met ieder 2 polen (in de analoge versie b.v. RLC-netwerken, Fant, 1968; Flanagan, 1972) kan ook nauw aansluiten bij een, in de fonetiek gebruikelijke beschrijving van het korte-termijn energiespectrum in termen van formanten. We hebben in par. 1.1 gezien dat frekwenties waar de omhullende in het energiespectrum lokaal maximaal is formanten kunnen vormen en dat die formanten ook belangrijke kenmerken kunnen zijn bij de perceptie van spraakklanken. Afzonderlijke formanten kunnen met tweede-orde, resonerende, deelfilters worden beschreven en daarmee wordt dan een samenhang gelegd tussen de parameters van zo'n deelfilter en afzonderlijke kenmerken in de spraakperceptie. Dat was een van de eisen die we aan het analyse-resynthesesysteem hebben gesteld.

Daarnaast hebben we als eis gesteld dat de parameters snel en rechtstreeks automatisch uit het spraaksignaal zelf berekend moeten kunnen worden. We zullen in het volgende hoofdstuk laten zien dat een digitaal hogere-orde filter volgens de techniek van 'invers filteren' (Markel, 1972) of 'linear predictive coding' LPC (Atal en Hanauer, 1971; Makhoul, 1975; Markel en Gray, 1976) automatisch berekend kan worden uit de gedigitaliseerde golfvorm van het spraaksignaal. Het in ons systeem toegepaste filter is een kombinatie van beide beschrijvingen.

1.3. HET DOOR ONS TOEGEPASTE MODEL

Wij zullen in ons analyse-resynthesesysteem de fysica van de spraakproduktie beschrijven met een digitaal bron-filtermodel waarvan het bronsignaal bestaat uit een periodieke impuls of witte ruis en het filter uit een cascade van 2e orde filters met louter polen, zoals is weergegeven in fig. 1.7.

Het model is digitaal omdat daarmee hoge precisie en grote flexibiliteit worden gekombineerd met eenvoud bij de implementatie in een digitale rekenmachine. Het filter bevat louter polen omdat daardoor de parameters op snelle en eenvoudige wijze uit het oorspronkelijke, gedigitaliseerde spraaksignaal zelf berekend kunnen worden. Het model bestaat uit een set van tweede orde deelfilters omdat de parameters daarvan nauw gerelateerd zijn aan afzonderlijke formanten. Het aantal deelfilters zal in principe vijf zijn, omdat in het frekwentiegebied tot 5 kHz doorgaans niet meer dan vijf formanten voorkomen en dus in

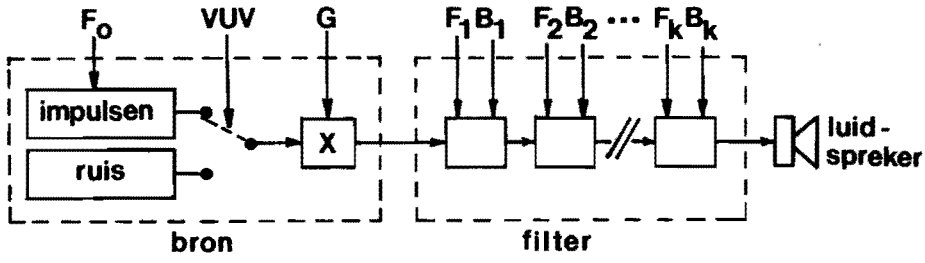

Fig.1.7. Het door ons toegepaste bron-filtermodel voor de produktie van spraakgeluid. Het filter is samengesteld uit een cascade van 2e orde deelfilters. De parameters van ieder filter zijn hier symbolisch aangeduid met afstemfrekwenties /F_k/ en bandbreedtes /B_k/.

totaal niet meer dan tien filterparameters nodig zijn. Verder hebben we in ons model gekozen voor een cascade van deelfilters omdat dan niet voor ieder filter naast de beide coëfficiënten ook nog een amplitudefactor hoeft te worden gespecificeerd. Een parallelschakeling (Holmes, 1973) zou dat wel vereisen.

Bron en filters zijn bestuurbaar; hun parameters veranderen op discrete punten in de tijd en karakteriseren dan de eigenschappen van het model gedurende een bepaalde periode. Een bij elkaar horend stel parameters, hier verder frame genoemd, beschrijft een eenvoudig lineair model dat een eerste benadering is van de fysica van de menselijke spraakproduktie. Dit model vertoont, naast overeenkomsten met de menselijke spraakproduktie, mede door zijn eenvoud, ook verschillen daarmee. De belangrijkste daarvan zullen we in de volgende paragraaf bespreken.

1.4. VERGELIJKING TUSSEN MODEL EN MENSELIJKE SPRAAKPRODUKTIE

Op een viertal punten kunnen we verschillen signaleren tussen model en werkelijke fysica van de menseljke spraakproduktie.

Het eerste verschil tussen model en menselijke spraak is dat het model niet voorziet in een gelijktijdige kombinatie van periodiek- en ruisbron, terwijl in werkelijkheid wel degelijk zulke combinaties voor kunnen komen, bv bij stemhebbende wrijfklanken.

Het is in principe mogelijk om zo'n kombinatie wel in het model op te nemen, bv door periodiek- en ruisbron hun eigen versterkingsfaktor mee te geven en dan beide bronsignalen bij elkaar te voegen aan de ingang van het filter. Daarmee blijken echter nog geen perceptief goede stemhebbende wrijfklanken te ontstaan. In natuurlijke spraak

zijn deze klanken niet zonder meer de som van een periodiek en een een ruisig signaal. Vaak is de periodieke component beperkt tot de lage en de ruis tot de hogere frekwenties, of is de ruis slechts in een deel van de grondtoonperiode aanwezig. Daarvan hebben we in fig. 1.4 een voorbeeld gezien voor de z.

Een principieel betere methode is door Makhoul ea (1978) voorgesteld. Daarbij wordt het frekwentiegebied in tweeën verdeeld en is het brongeluid beneden een bepaalde grensfrekwentie periodiek en daarboven ruis. Deze werkwijze is door Darwin (1982) geïmplementeerd maar levert perceptief slechts een geringe verbetering op, die in geen verhouding staat tot de ingewikkelde en veel rekentijd vergende bepaling van de grensfrekwentie en de resynthese van de spraak.

We zien hier dan ook af van een mengvorm van periodiek en ruisig brongeluid en beperken ons tot een binaire keuze tussen beide. In hoofdstuk 4 en 5 zullen we nagaan wat de gevolgen van deze beperking zijn. Tevens zal dan blijken dat perceptief goede stemhebbende wrijfklanken kunnen worden geproduceerd door periodiek- en ruisbron in de tijd af te wisselen.

Als tweede verschil tussen model en menselijke spraakproduktie merken we op dat het model filters bevat waarvan de overdrachtsfunktie uitsluitend polen heeft en dus niet de nulpunten beschrijft die theoretisch voor kunnen komen in het spectrum van nasale klanken, waarbij de mondopening is afgesloten. In principe zou een nulpunt weliswaar door een groot aantal polen willekeurig goed kunnen worden benaderd, maar dat heeft voor ons niet zoveel betekenis omdat we het filter willen realiseren met een klein aantal parameters. Op de gevolgen van deze beperking komen we uitvoerig terug in hoofdstuk 4 en 5.

Een derde beperking van het model is dat bron en filter in het model onafhankelijk zijn en elkaar niet belasten. Bij menselijke spraakproduktie is dat niet het geval; de akoestische impedantie van de stembanden enerzijds en de ligging en bandbreedte van vooral de lagere formanten anderzijds beïnvloeden elkaar enigszins (Flanagan, 1972). Hoewel deze interaktie via de modelparameters nog ingebouwd zou kunnen worden, maakt dat het modelaanzienlijk ingewikkelder en we zien daar in ons geval dan ook van af.

Als laatste memoreren we nog dat in het model het spraaksignaal wordt beschreven als stapsgwijze opeenvolging van lokaal stationaire signalen. De parameters van de opeenvolgende frames beschrijven het model op discrete tijdstippen, en blijven dan geldig voor een bepaalde tijdsduur, de frameperiode. Binnen die frameperiode worden ze konstant verondersteld. In de werkelijke spraakproduktie veranderen de eigen-

schappen van het mondkanaal echter continu. Maar de snelheid waarmee dat gebeurt is beperkt en een beschrijving in discrete stappen is dan adequaat, mits die stapgrootte of frameperiode voldoende klein is. Alleen om praktische redenen van zuinigheid moet de frameperiode niet onnodig klein zijn. Op wat 'voldoende groot' en 'onnodig klein' is komen we nader terug in hoofdstuk 4 en 5.

In de volgende hoofdstukken zullen we laten zien dat dit eenvoudige model, ondanks bovengenoemde beperkingen, zeer goed bruikbaar is voor de doeleinden die we in het vorige hoofdstuk hebben geformuleerd. We zullen nu behandelen hoe we de parameters van het model kunnen berekenen uit het spraaksignaal zelf: de analyse.

2 ANALYSE VAN HET SPRAAKSIGNAAL

In dit hoofdstuk wordt uiteengezet hoe de modelparameters als funktie van de tijd bepaald kunnen worden uit het gedigitaliseerde spraaksignaal zelf. Volgens de uit de literatuur bekende techniek van invers filteren of lineaire predictie (LPC) wordt voor een gegeven spraaksegment een M^e orde analysefilter bepaald. De coëfficiënten daarvan, de zgn a-parameters, worden in het tijddomein zó berekend dat de totale energie aan de uitgang van het filter minimaal is voor het gegeven spraaksegment. In het frekwentiedomein is dan, bij het gegeven aantal van M filtercoëfficiënten, het energiespectrum van het ingangssignaal zo goed mogelijk vlak gestreken. Geïnverteerd levert dat analysefilter dan een zo goed mogelijke benadering van de omhullende van het energiespectrum van het ingangssignaal.

Aan deze a-parameters kan echter niet rechtstreeks spectrale informatie worden ontleend over het analysefilter. Dat kan wel als we het M^e orde filter opvatten als een cascade van M/2 tweede-orde filters en de coëfficiënten van deze tweede-orde secties, de zgn pq-parameters afsplitsen uit de berekende a-parameters. Door vervolgens deze pq-paren om te rekenen naar afstemfrekwenties F en kwaliteitsfaktoren Q, de zgn FQ-paren, kunnen daarmee anti-resonanties (dalen) van het analysefilter worden geassocieerd. Ieder FQ-paar is dan te associëren met een resonantie (top) in het ingangsspectrum. Naar analogie van een spectrogram vormen de FQ-parameters, uitgezet als funktie van de tijd, het "antiresonantiediagram" van het analysefilter. Dit diagram, hier verder afgekort tot resogram, speelt een grote rol bij de presentatie van de analyseresultaten.

In de tweede paragraaf wordt besproken hoe de bronparameters van het model worden berekend. De amplitudeversterkingsfaktor G wordt berekend uit de gevonden filtercoëfficiënten en de autocorrelaties van het inganssignaal. We zullen dat in hoofdstuk 3 aantonen; pas bij de synthese zijn de voorwaarden te formuleren waaruit de energie en dus G kan worden bepaald. Ter bepaling van de stem/stemloosparameter VUV wordt nagegaan of het spraaksegment voldoende periodiek is. Als dat het geval is, wordt vervolgens de periode van de best passende grondtoon F_0 berekend met behulp van de door Duifhuis, Willems en Sluyter (1982) ontwikkelde methode van de harmonische zeef, die is gebaseerd op Goldstein's (1973) theorie voor de menselijke toonhoogteperceptie.

Vervolgens geven we in de derde paragraaf van dit hoofdstuk een beschrijving van de implementatie van het analyseproces op de computer, gevolgd door de presentatie en bespreking van een aantal analyseresultaten in de vorm van resogrammen. Die resultaten bestaan

echter nog uit ongeordende paren coëfficiënten die 2e orde deelfilters beschrijven waarvan de overdrachtsfunktie ook reele poolparen kan bevatten. Laatstgenoemde zijn niet in het resogram vertegenwoordigd. Daarom wordt tot besluit in de laatste paragraaf behandeld hoe we de analyseresultaten zó kunnen bewerken dat steeds alle M/2 deelfilters resonerend zijn en hun afstemfrekwenties geordend kunnen worden langs de frekwentieas en ze ook in een resogram zijn weer te geven.

2.1. BEPALING VAN DE FILTERPARAMETERS

In het vorige hoofdstuk is een eenvoudig model beschreven voor de fysica van de spraakproduktie, waarin een spectraal wit bronsignaal een variabel filter exciteert. De overdrachtsfunktie van dit produktiefilter van het model bepaalt de spectrale omhullende van het uitgangssignaal en dus de omzetting van het witte, vlakke bronspectrum naar het gekleurde, gepiekte spectrum aan de uitgang van het model.
Het uitgangspunt dat het bronspectrum wit is levert ons nu de mogelijkheid om voor een bepaald stukje spraak, via de techniek van invers filteren (Markel, 1972), de parameters van het filter rechtstreeks uit het spraaksignaal zelf te bepalen. Stel dat we een spraaksegment met een analysefilter zodanig filteren dat aan de uitgang van dit filter het spectrum een vlakke omhullende heeft, dus spectraal wit is. Dan moet de overdrachtsfunktie van dat analysefilter de geïnverteerde zijn van het produktiefilter dat we zoeken, immers dit laatste heeft een wit ingangsspectrum. Dat betekent dat we de parameters van het produktiefilter kunnen vinden door de overdrachtsfunktie van het analysefilter te inverteren; analyse- en produktiefilter zijn elkaars inversen.
Hoe we de parameters van het analysefilter zó berekenen dat het spectrum aan de uitgang ervan vlak wordt, is het onderwerp van de volgende paragraaf.

2.1.1. Bepaling van de a-parameters van het analysefilter

We nemen aan dat het analysefilter A fysisch realiseerbaar en dus causaal is en dat het lineair en tijdonafhankelijk is voor een bepaalde, nog nader te specificeren tijdsduur. Verder dat het een digitaal filter is, waarvan de overdrachtsfunktie louter nulpunten heeft (bv. Rabiner en Schafer, 1978). Voor zo'n filter wordt in het tijddomein het uitgangssignaal e_n (het uitgangssample op tijdstip nT, met T de bemonsteringsperiode en n geheel) alleen bepaald door het ingangssignaal s_n op datzelfde tijdstip en een lineaire kombinatie van M voorafgaande ingangssamples:

$$(2.1.1) \qquad e_n = s_n + \sum_{k=1}^{M} a_k\, s_{n-k} = \sum_{k=0}^{M} a_k\, s_{n-k}, \qquad \text{met } a_0 = 1.$$

Het uitgangssignaal e_n is dus de convolutie van het ingangssignaal s_n en de impulsresponsie $/a_k/$. Alleen de M voorafgaande samples uit het verleden dragen, elk voorzien van een eigen weegfaktor a_k bij tot het uitgangssample e_n, zie fig. 2.1. De orde M van het filter geeft het aantal coëfficiënten aan, dus het tijdsbereik waarover het filter de

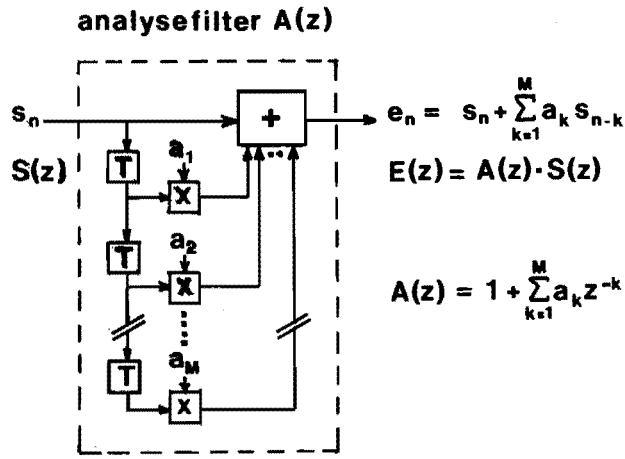

Fig.2.1. Digitaal Me orde analysefilter. Het uitgangssample e_n op tijdstip nT wordt gegeven door de som van het ingangssample s_n op datzelfde tijdstip en M gewogen daaraan voorafgaande ingangssamples. De weegfaktoren /a_k/ worden bij de analyse zó berekend dat de energie van het uitgangssignaal minimaal is.

---

samples uit het verleden onthoudt.

Voor de (totale) energie E van het uitgangssignaal, gedefinieerd door:

(2.1.2) $$E = \sum_n e_n^2 , \quad n = -\infty ,..+\infty$$

geldt dan met (2.1.1):

$$E = \sum_n (\sum_{k=0}^{M} a_k \, s_{n-k})^2$$

of

$$E = \sum_{i=0}^{M} a_i \sum_{k=0}^{M} a_k \sum_n s_{n-i} \, s_{n-k},$$

of:

(2.1.3) $$E = \sum_{i=0}^{M} a_i \sum_{k=0}^{M} a_k \, R_{i-k},$$

waarin:

(2.1.4) $$R_{i-k} = \sum_n s_{n-i} \, s_{n-k}$$

de $(i-k)^e$ autocorrelatie van het ingangssignaal s_n definieert. Hoewel n in principe van $-\infty$ tot $+\infty$ loopt, zijn in ons geval per definitie alle ingangssamples /s_n/ buiten het analysevenster, dat uit N samples bestaat, nul. Voor een minimale energie E is dan, na partiële differentiatie van (2.1.3) naar de filtercoëfficiënten /a_k/, af te leiden

dat als voorwaarde daarvoor geldt:

$$\sum_{k=0}^{M} a_k R_{i-k} = 0 , \qquad i = 1...M$$

of:

$$\sum_{k=1}^{M} a_k R_{i-k} = - R_i , \qquad i = 1...M \qquad (2.1.5)$$

omdat a_o per definitie 1 is.
Dit stelsel (2.1.5) van M vergelijkigen met de M filtercoëfficiënten $/a_k/$ als onbekenden kan snel recursief worden opgelost (Muller, 1975), na berekening van de autocorrelaties $/R_{i-k}/$ van het ingangssignaal volgens (2.1.4).
Wanneer de filtercoëfficiënten aan (2.1.5) voldoen is de energie E van het uitgangssignaal minimaal en deze wordt dan gegeven door:

$$E_m = \sum_{i=0}^{M} a_i R_i. \qquad (2.1.6)$$

Bij het berekenen van de filtercoëfficiënten $/a_k/$ wordt alleen gebruik gemaakt van de autocorrelaties (2.1.4) van het ingangssignaal. Dat signaal zelf is daarbij niet nader gespecificeerd en mag dus ook ruis zijn. Resultaat is steeds een set filtercoëfficiënten waarvoor geldt dat zij een filter van de gegeven orde M definiëren dat de energie van het uitgangssignaal voor het gegeven ingangssignaal zo klein mogelijk maakt. Hoe groot deze minimale energie E_m van het signaal aan de uitgang, het 'restsignaal', dan is, hangt af van M. Naarmate het filter meer coëfficiënten heeft zal het beter in staat zijn zijn taak te vervullen en zal de energie van het restsignaal kleiner zijn. E_m neemt monotoon af met toenemende M (Atal en Hanauer, 1971).

In het frekwentiedomein wordt het analysefilter gekarakteriseerd door zijn (complexe) overdrachtsfunktie A(z), de z-getransformeerde van de impulsresponsie $/a_k/$ van het filter, dus:

$$A(z) = \sum_{k=0}^{M} a_k z^{-k} . \qquad (2.1.7)$$

Hierin is $z = \exp(sT)$, met T de bemonsteringsperiode of reciproke samplefrekwentie en s de complexe hoekfrekwentie $s = \sigma + jw$, met als reëel deel σ en als imaginair deel de hoekfrekwentie $w = 2\pi f$.
De z-getransformeerde E(z) van het uitgangssignaal e_n wordt dan gegeven door (fig. 2.1):

$$E(z) = A(z)\, S(z), \qquad (2.1.8)$$

waarin S(z) de z-getransformeerde is van het ingangssignaal s_n.
Voor de energie E van het uitgangssignaal geldt dat de kwadratensom (2.1.2) in het tijddomein ook geschreven kan worden als integratie in

het frekwentiedomein (theorema van Parseval):

$$(2.1.9) \qquad E = \sum_n e_n 2 = (1/2\pi) \int_{-\pi}^{+\pi} |E(w)|\,2dw$$

Minimaliseren van E volgens (2.1.5) betekent dat, geïntegreerd over het gehele frekwentiegebied, de energie van het uitgangssignaal, bij gegeven M, zo klein mogelijk wordt gemaakt. Voor het betreffende spraaksegment van N samples in het analysevenster is de spectrale omhullende van het uitgangssignaal, door het filter waarvan de coëfficiënten aan (2.1.5) voldoen, dan zo vlak mogelijk gemaakt (Markel en Gray, 1976).

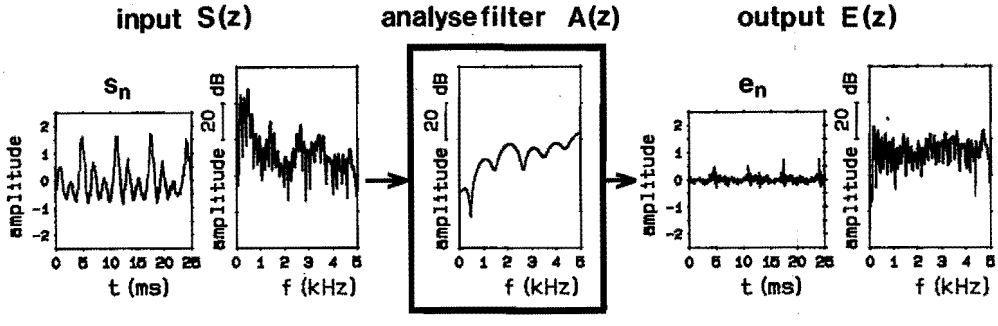

Fig.2.2. Energiespectrum (midden) van een 10e orde analysefilter A(z), berekend voor een periodiek (stemhebbend) ingangssignaal. Het spectrum van het restsignaal aan de uitgang van het filter is bij benadering vlak (wit) geworden.

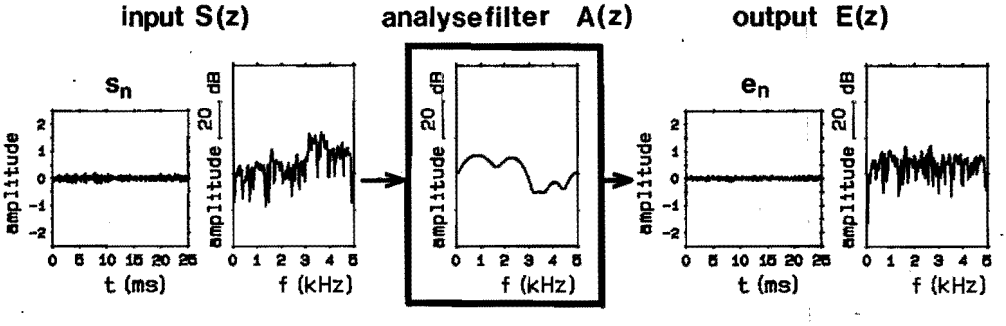

Fig.2.3. Als Fig. 2.2 maar nu voor een ruisig (stemloos) ingangssignaal.

Een voorbeeld van een 10^e orde filter berekend voor zowel een periodiek als een ruisig ingangssignaal, is weergegeven in fig. 2.2 en 2.3. Links staan golfvorm en energiespectrum van beide ingangssignalen, rechts die aan de uitgang. We zien hoe het energiespectrum van het restsignaal e_n door het filter is vlakgestreken; de pieken (resonanties) in de omhullende van het ingangsspectrum zijn door de dalen (antiresonanties) van het berekende analysefilter 'geneutraliseerd'. Het daarvoor benodigde energiespectrum van het filter A(z) is eveneens in fig. 2.2 en 2.3 weergegeven.

In principe zouden we nu voor de representatie van het analysefilter kunnen volstaan met de M coëfficiënten $/a_k/$. Immers zij vormen de impulsresponsie van het filter, waarmee de overdrachtsfunktie A(z), en via de Fouriertransformatie dus ook het energiespectrum, volledig is bepaald. Deze a-parameters hebben echter het nadeel dat ze weinig direkt inzicht verschaffen omtrent ligging en vorm van de spectrale dalen in het analysefilter en dus ook niet over de pieken in het ingangsspectrum. Dat inzicht kan wel verkregen worden als we de a-parameters omrekenen naar een produkt van coëfficiënten van 2^e orde filters. In de volgende paragrafen zullen we laten zien hoe aan deze pq-parameters wel nadere spectrale gegevens kunnen worden ontleend.

2.1.2. Bepaling van de pq-parameters van het analysefilter

We hebben de coëfficiënten $/a_k/$ berekend van een M^e orde filter dat de spectrale omhullende van het ingangssignaal zo goed mogelijk vlak strijkt. Pieken (resonanties) worden door het filter met dalen (antiresonanties) naar vermogen geneutraliseerd. Om het verband tussen filterparameters en spectrale (anti)resonanties expliciet te maken is het nodig dit M^e orde analysefilter om te rekenen naar een cascade van $M/2$ 2^e orde filters. Eén zo'n 2^e orde sectie kan in principe één resonantiepiek in het ingangsspectrum voor zijn rekening nemen. Het M^e orde filter heeft als overdrachtsfunktie

$$A(z) = 1 + \sum_{k=1}^{M} a_k \, z^{-k} \quad , \tag{2.1.8}$$

en dit polynoom is ook als produkt van kwadratische termen te schrijven. Aannemende dat M even is geldt:

$$A(z) = \prod_{k=1}^{M/2} (1 + p_k \, z^{-1} + q_k \, z^{-2}) = \prod_{k=1}^{M/2} D_k(z) \quad , \tag{2.1.20}$$

waarin $/p_k,q_k/$ de filtercoëfficiënten zijn van de 2^e orde secties, die

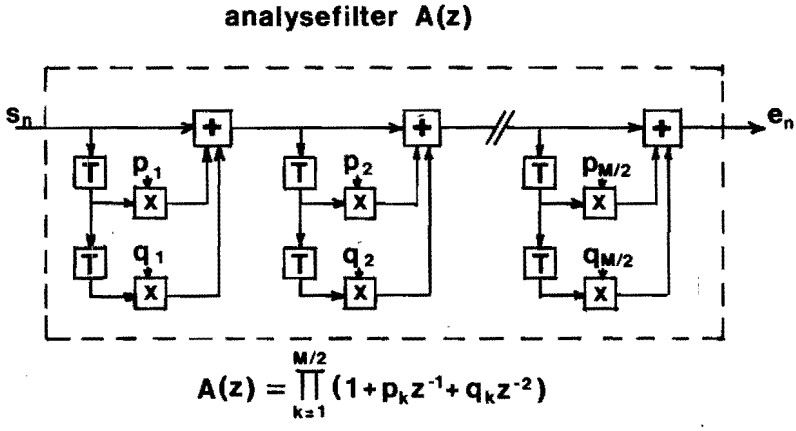

Fig.2.4. Het analysefilter van Fig. 2.1, nu in de vorm van een cascade van M/2 2e orde deelfilters met als parameters de coëfficiënten $/p_k, q_k/$.

---

we verder pq-parameters zullen noemen, zie fig. 2.4.

Om die kwadratische termen van een hogere-orde polynoom af te splitsen kan de Bairstow-methode worden gebruikt (Fröberg, 1969, Muller, 1975), een numeriek-iteratieve methode om nulpunten van een polynoom te bepalen. Daarmee is dan het analysefilter A(z) omgerekend naar een cascade van 2^e orde deelfilters $/D_k(z)/$. Om het verband tussen spectrale resonanties en filtercoëfficiënten te leggen zullen we nu eerst enkele eigenschappen van zo'n 2^e orde filter bespreken.

Elke sectie wordt gekarakteriseerd door zijn overdrachtsfunktie:

(2.1.21) $D(z) = 1 + pz^{-1} + qz^{-2}$,
of $D(z) = z^{-2}\,(z-z_1)(z-z_2)$.

In het complexe z-vlak heeft zo'n sectie dus behalve een dubbele pool in de oorsprong twee nulpunten, z_1 en z_2, die reëel of toegevoegd complex kunnen zijn, immers de oorspronkelijke coëfficiënten $/a_k/$ van het A-polynoom zijn alle reëel. Verder kan worden aangetoond (Markel en Gray, 1976) dat de wortels van het A-polynoom, die gegeven zijn door A(z) = 0, alle binnen de eenheidscirkel $|z| = 1$ liggen. Dus ook de nulpunten z_1 en z_2 van de afzonderlijke 2^e orde polynomen voldoen aan die voorwaarde.
De nulpunten van (2.1.21) worden gegeven door de wortels van D(z) = 0, dus:
(2.1.22) $z_{1,2} = -p/2 \pm j\sqrt{q - p^2/4}$.

We kunnen nu 3 gevallen onderscheiden (fig. 2.5):

- z_1 en z_2 zijn toegevoegd complex. Dit is het geval als in (2.1.22) $q > p^2/4$. Voor de moduli geldt dan:
  (2.1.23) $|z_1| = |z_2| = \sqrt{q}$
  en voor de argumenten:
  (2.1.24) $\arg(z_1) = -\arg(z_2) = \arccos(-p/2\sqrt{q})$.
  Uit de voorwaarde dat de nulpunten altijd binnen de eenheidscirkel liggen volgt dan nog dat $|q| < 1$. In het (reële) pq-vlak uitgezet zien we in fig. 2.5b dat toegevoegd complexe nulpunten optreden wanneer de pq-paren liggen binnen het segment gevormd door de parabool $p^2 = 4q$ en de rechte $q = 1$. Dus alleen filters waarvan de coëfficiënten binnen dit gebied liggen kunnen met een resonantie in het ingangsspectrum worden geassocieerd.
- z_1 en z_2 reëel en gelijk. Dit is het geval als $q = p^2/4$, dus als: $|z_1| = |z_2| = p/2$ en $\arg(z_1) = \arg(z_2) = \pi$.
  Uit de voorwaarde $|z| < 1$ volgt dan nog dat $|p| < 2$, dus de parameters p en q liggen in dit geval in fig. 2.5b op de parabool $p^2 = 4q$ tussen $p = -2$ en $p = +2$.
- z_1 en z_2 reëel maar ongelijk. In dit geval geldt:
  (2.1.25) $z_{1,2} = -p/2 \pm \sqrt{p^2/4 - q}$.
  Ook nu moet voldaan worden aan de voorwaarde dat $|z| < 1$, dus ook:
  $$\sqrt{p^2/4 - q} < 1 - |-p/2| ,$$
  hetgeen na kwadrateren oplevert:
  $$p^2/4 - q < 1 + p^2/4 - |p| ,$$
  zodat moet gelden:
  (2.1.26) $q > p - 1$ voor $p > 0$ en
  (2.1.27) $q > -p - 1$ voor $p < 0$.
  In fig. 2.5b liggen de parameters p en q in dit geval dus binnen de driehoek gevormd door de rechten $q = -p-1$, $q = p-1$ en $q = 1$, maar onder de parabool $p^2 = 4q$.

Samenvattend: een 2^e orde filter dat afgesplitst is van het hogere orde analysefilter kan als coëfficiënten slechts pq-combinaties opleveren die binnen de driehoek in fig. 2.5b liggen. Als de pq-paren daarbij op of onder de parabool $p^2 = 4q$ liggen, zijn de nulpunten van de overdrachtsfunktie reëel. De sectie draagt er dan toe bij om de spectrale omhullende van het ingangssignaal vlak te strijken, maar kan niet worden geassocieerd met een afzonderlijke resonantiepiek in dat spectrum. Alleen als pq-paren boven de parabool liggen heeft zo'n sectie toegevoegd complexe nulpunten, vormt dan een anti-resonantie en is te associëren met een resonantie in het ingangsspectrum.

In de volgende paragraaf zal het verband worden afgeleid tussen de pq-parameters van zo'n 'resonerend' digitaal deelfilter en de daarbij behorende afstemfrekwentie en bandbreedte van het analoge filter.

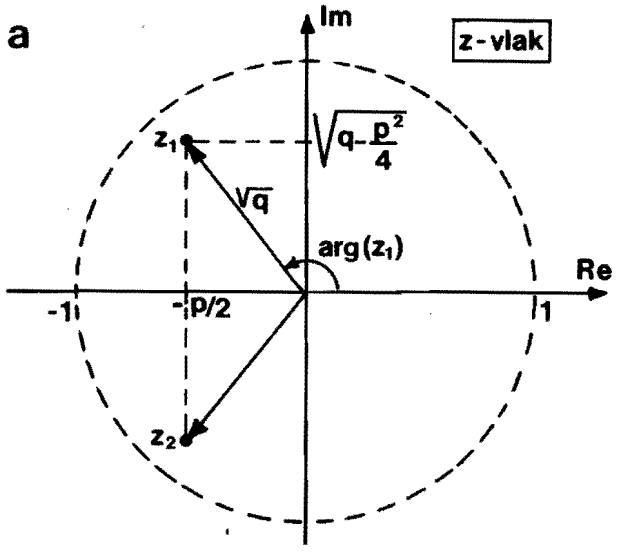

Fig.2.5a. De nulpunten z_1 en z_2 in de complexe overdrachtsfunktie van een digitaal 2e orde deelfilter. Voor een stabiel filter liggen deze binnen de eenheidscirkel $|z| = 1$.

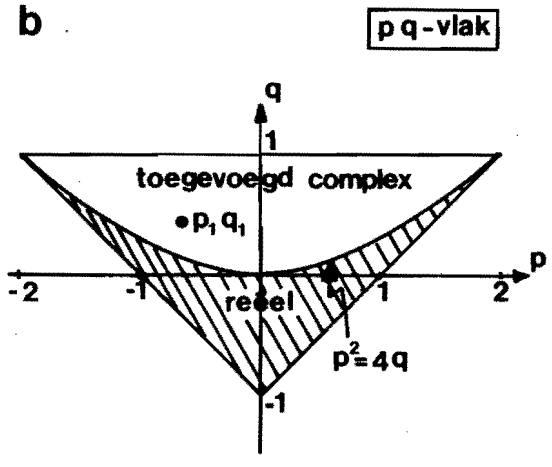

Fig.2.5b. De pq-parameters van een 2e orde deelfilter. Voor een stabiel filter liggen de pq-combinaties binnen de driehoek; als ze tevens boven de parabool $p^2 = 4q$ liggen zijn de nulpunten z_1 en z_2 toegevoegd complex en is het deelfilter resonerend.

2.1.3. Relatie tussen digitale pq- en analoge FB-parameters

Analoog zijn 2^e orde netwerken ('kringen') zoals RLC-netwerken te karakteriseren door 2 grootheden: de (fase)resonantiefrekwentie w_o en de kwaliteitsfaktor Q, die een maat is voor de selectiviteit van de kring.
Voor een seriekring zijn deze resp. gedefinieerd door:

(2.1.30) $w_o = 1/\sqrt{LC}$

en

(2.1.31) $Q = w_oL/R = 1/w_oRC$.

In het complexe s-vlak (met $s = \sigma + jw$, en $w = 2\pi f$ de hoekfrekwentie) wordt de overdrachtsfunktie van zo'n seriekring (anti-resonantie) gegeven door b.v.:

$$H(s) = s^2 + (w_o/Q)\, s + w_o^2 \quad ,$$

of

$$H(s) = (s - s_1)(s - s_2) \quad ,$$

waarin dus de nulpunten s_1 en s_2 worden gegeven door:

(2.1.32) $s_{1,2} = -w_o/2Q \pm jw_o\sqrt{1-1/4Q^2}$.

De ligging van deze nulpunten in het complexe s-vlak is geschetst in fig. 2.6, voor het geval dat zij toegevoegd complex zijn, dus $Q > 1/2$.

Het imaginaire deel w_1 van de nulpunten heet ook de eigenfrekwentie en is gegeven door:

(2.1.33) $w_1 = w_o\sqrt{1-1/4Q^2}$.

Tenslotte wordt nog de amplituderesonantiefrekwentie w_a onderscheiden, gedefinieerd door:

(2.1.34) $w_a = w_0\sqrt{1-1/2Q^2}$,

dat is de hoekfrekwentie waarbij de absolute waarde van de overdrachtfunktie minimaal is. Deze w_a zal hier verder niet meer worden gebruikt. Voor grote kwaliteitsfaktor Q vallen zowel w_a als w_1 vrijwel samen met de resonantiefrekwentie w_o.

In plaats van de eigenfrekwentie w_1 (in rad/s) wordt hierna veel gebruik gemaakt van de afstemfrekwentie F (in Hz), gedefinieerd door:

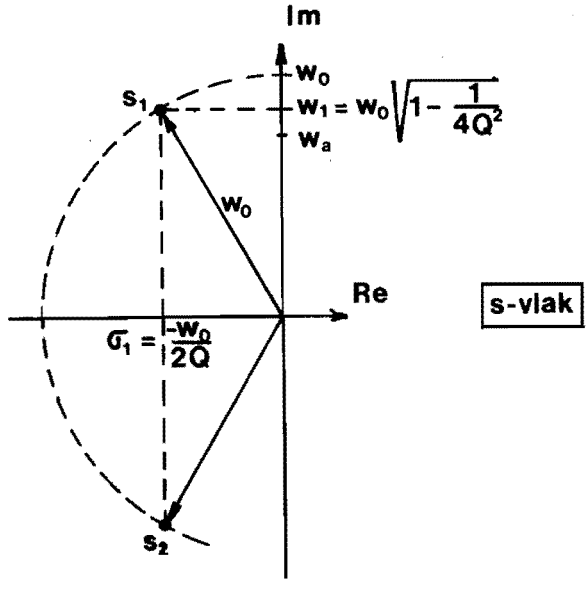

Fig.2.6. De nulpunten s_1 en s_2 in de complexe overdrachtsfunktie van een analoog 2e orde filter.

(2.1.35) $F = w_1/2\pi$.

De bandbreedte Δw (in rad/s) van de kring is gedefinieerd door $\Delta w = w_o/Q$, waarbij dus de relatieve bandbreedte $\Delta w/w_o$ de reciproke kwaliteitsfaktor is. De bandbreedte B (in Hz) is dan (fig. 2.6) gegeven door $\Delta w = 2\pi B = w_o/Q = -2\sigma_1$, of:

(2.1.36) $B = -\sigma_1/\pi$.

De uitdrukkingen (2.1.35) en 2.1.36) voor resp. afstemfrekwentie en bandbreedte van het filter zijn alleen gedefinieerd als de nulpunten toegevoegd complex zijn, dus als w_1 positief is.
Voor grote kwaliteitsfaktor Q volgt uit (2.1.33) met $Q^2 >> 1/4$ dat $w_o \approx w_1$, zodat dan de kwaliteitsfaktor Q ook te schrijven is als $Q = w_o/\Delta w = w_1/2\pi B$, of:

(2.1.31a) $Q = F/B$.

Deze benadering wordt hier verder gehanteerd als definitie van de kwaliteitsfaktor.

Het verband tussen (analoge) afstemfrekwentie F en bandbreedte B en (digitale) coëfficiënten p en q van het overeenkomstige 2^e orde filter wordt nu gegeven door de relatie tussen het s-vlak en het z-vlak. Die is via de z-transformatie gegeven door:

(2.1.37) $\quad z = \exp(sT) = \exp(\sigma T + jwT)$,

waarin T weer de bemonsteringsperiode is. Voor het geval van toegevoegd complexe nulpunten volgt uit (2.1.35) en (2.1.36) voor de natuurlijke logaritme van (2.1.37):

(2.1.38) $\quad \ln|z| = \sigma T = -\pi BT$

en

(2.1.39) $\quad \arg(z) = wT = 2\pi FT$.

Voor de toegevoegd complexe z_1 en z_2 is dan met (2.1.23) en (2.1.24) hieruit af te leiden dat:

(2.1.40) $\quad F = (1/2\pi T) \arccos(-p/\sqrt{2q})$, met $q > p^2/4$

en

(2.1.41) $\quad B = (-1/\pi T) \ln \sqrt{q}$, met $q > 0$

en omgekeerd:

(2.1.42) $\quad q = \exp(-2\pi BT)$

en

(2.1.43) $\quad p = -2\sqrt{q} \cos(2\pi FT)$.

Hiermee zijn dan de relaties gegeven tussen de afstemfrekwentie F en bandbreedte B of kwaliteitsfaktor Q enerzijds en de digitale 2^e orde filtercoëfficiënten p en q anderzijds. Ook nu geldt natuurlijk dat de afstemfrekwentie alleen is gedefinieerd door (2.1.40) als de arccos bestaat, dus als $q > p^2/4$, en de pq-combinatie dus binnen het paraboolsegment in fig. 2.5 ligt.

We hebben nu de mogelijkheid gekregen om het M^e orde analysefilter weer te geven met (ten hoogste) M/2 antiresonanties, in de vorm van FB- of FQ-parameters, met F de afstemfrekwentie, B de bandbreedte en Q = F/B de kwaliteitsfaktor van de afzonderlijke antiresonanties. Deze representatie met FQ-paren zal hierna veel worden gebruikt om de analyseresultaten in een resogram visueel weer te geven, voor zover de paren althans gedefinieerd zijn. Voorbeelden daarvan zullen we in par. 2.4 geven. Wat we doen met niet-resonerende deelsecties komt aan de orde in par. 2.5. Eerst zullen we in de volgende paragraaf behandelen hoe we de nog resterende modelparameters bepalen.

2.2. BEPALING VAN DE BRONPARAMETERS

2.2.1 Bepaling van de amplitude G

De amplitudeversterkingsfaktor G wordt bij de analyse berekend uit het inwendig produkt van de gevonden filtercoëfficiënten $/a_k/$ en de autocorrelaties $/R_k/$ van het signaal in het analysevenster. Dit inprodukt (2.1.6) is gelijk is aan de energie E_m van het restsignaal, de nog resterende energie aan de uitgang van het analysefilter. We zullen in hoofdstuk 3 aantonen dat met $G^2 = E_m$ het spraaksegment van het betreffende frame na resynthese de juiste energie heeft.

Tot nu toe is bij de analyse geen onderscheid gemaakt tussen ingangsignalen die periodiek of ruisig zijn. Dat hoefde ook niet, omdat in beide gevallen de analyse identiek is, ze berust op het bepalen van de eerste M autocorrelaties van het ingangssignaal. In beide gevallen is de omhullende van het uitgangsspectrum zo vlak mogelijk gemaakt, onafhankelijk van de vraag of de fijnstruktuur periodiek of ruisig is.

Toch willen we graag, conform het model van hoofdstuk 1, onderscheid maken tussen deze twee bronsignalen omdat we straks bij de resynthese een van beide zullen moeten kiezen. De vraag die we nu aan de orde stellen is dan ook: is het ingangssignaal ruisig of periodiek, en als het periodiek is, wat is dan de frekwentie van de grondtoon?

2.2.2. Bepaling stem/stemloos parameter VUV

We hebben in hoofdstuk 1 gezien dat in de menselijke spraakproduktie alleen bij stemhebbende klanken de -12 dB/oct helling van het spectrum van de stembandtrillingen meespeelt. Samen met de +6 dB/oct van het stralingseffekt aan de mondopening resulteert dit bij stemhebbende klanken in een helling van -6 dB/oct, terwijl bij stemloze klanken alleen de +6 dB/oct van de straling optreedt. Een voor de handliggende stem/stemloos detector is dan ook de bepaling van de globale helling van de omhullende van het energiespectrum van het spraaksignaal. Dat kan gebeuren door berekening van de eerste (genormeerde) autocorrelatie van het ingangssignaal, gedefinieerd door:

$$(2.2.1) \qquad R_1/R_0 = \sum_n s_n\, s_{n-1} \;/\; \sum_n s_n^2$$

Dit quotient is op te vatten als de coëfficiënt van een analysefilter met orde M = 1, immers dan wordt (2.1.5):

$$a_1 = - R_1/R_0$$

en a_1 is dus de coëfficiënt van een 1^e orde filter dat het energie-

spectrum zo vlak mogelijk maakt. Voor stemhebbende ingangssignalen is R_1/R_0 bijna 1; er is een hoge correlatie tussen twee opeenvolgende samples, terwijl bij stemloze ingangssignalen deze correlatie klein of negatief is. Voorbeelden van dergelijke spectra hebben we gezien in hoofdstuk 1.

In ons analysesysteem baseren we de stem/stemloosklassificatie op twee elementen: de verhouding R_1/R_0 en de waarde van R_0 zelf, dus de totale energie in het analysevenster. Als die energie hoog is hebben we vaak met klinkers te doen. Het signaal wordt bij hoge R_0 dan ook als stemhebbend geklassificeerd, tenzij er zó weinig correlatie in het signaal is dat R_1/R_0 lager is dan 0.4. Omgekeerd is bij stemloze klanken de energie meestal relatief laag. Bij lage R_0 wordt het signaal dan ook als stemloos geklassificeerd, tenzij de verhouding R_1/R_0 hoger is dan 0.9.

Deze eenvoudige stem/stemloosdetector werkt uiteraard niet foutloos. Met name bij combinaties van periodieke en ruisige signalen, dus bij stemhebbende wrijfklanken kan achteraf corrigeren 'met de hand' betere resultaten opleveren. We komen daarop uitvoerig terug in hoofdstuk 5.

2.2.3 Bepaling van de grondtoonfrekwentie F_0

Wanneer het ingangssignaal als stemhebbend is beoordeeld is de volgende vraag die we moeten beantwoorden wat dan de herhalingsfrekwentie F_0 van de impuls van het bronsignaal is. Daarmee zijn we gekomen bij het probleem van toonhoogtemeting in spraak.

In de literatuur is een indrukwekkende hoeveelheid toonhoogtemeters en -algorithmen beschreven, waaraan regelmatig nieuwe procedures voor het meten van de toonhoogte in spraak worden toegevoegd. Geen van alle werkt foutloos; maar er zijn voor de praktijk zeer bruikbare methoden en de bekendste daarvan zijn onderling vergeleken in Rabiner et al (1976) en perceptief geevalueerd in McGonegal et al (1977).

In ons systeem is de door Duifhuis, Willems en Sluyter (1982) ontwikkelde toonhoogtemeter DWS toegepast, waarin de grondtoon wordt bepaald van het harmonisch spectrum dat het best past op afzonderlijke pieken in het spectrum van het spraaksignaal. Deze methode is gebaseerd op Goldsteins (1973, 1978) theorie voor de waarneming van toonhoogte van complexe tonen, dat zijn tonen die uit meer dan een harmonische bestaan. DWS is de enige toonhoogtemeter waaraan een een perceptieve theorie ten grondslag ligt en blijkt in de praktijk ook doorgaans de minste fouten te leveren. Voor details verwijzen we naar Duifhuis et al (1982); de teksten van de Fortran-procedures zijn gegeven in L.F.Willems (1982).

Hiermee is de analyse kompleet en in de volgende paragraaf zullen we bespreken hoe het analyseproces in de praktijk wordt uitgevoerd voor langere spraakuitingen.

2.3. PRAKTISCHE UITVOERING VAN DE ANALYSE

2.3.1. Voorbewerking en spraakinname

Voorafgaande aan de eigenlijke analyse in de computer moet het spraaksignaal, afkomstig van b.v. microfoon of magneetband, eerst van deze analoge vorm worden omgezet in een digitale, een reeks getallen. Daarbij wordt een procedure gevolgd die vrijwel standaard is.

De eerste stap daarin bestaat uit het filteren van het analoge spraaksignaal. Het spectrum wordt aan de hoge kant begrensd tot een frekwentie die de helft is van de samplefrekwentie waarmee we het signaal gaan bemonsteren. Deze begrenzing is nodig om ongewenste bijverschijnselen (vouwvervorming) bij het digitaliseren te vermijden. Bij een voor ons interessant frekwentiegebied van het te analyseren spraaksignaal van 0 tot 5 kHz worden dus alle hogere komponenten weggefilterd.

Aan de lage kant wordt de gelijkstroomkomponent geblokkeerd, dwz dat de gemiddelde waarde van het analoge spraaksignaal nul wordt gemaakt. Dit is niet strikt noodzakelijk voor de analyse, maar wordt gedaan omdat een hoog DC-nivo de uitstuurbaarheid van de analoog-digitaalomzetter nodeloos zou beperken. Bovendien beinvloedt zo'n komponent de waarde van de autocorrelaties en daarmee zowel de filtercoëfficiënten als de uitkomst van de stem/stemloos detektor.

De tweede stap is dan de eigenlijke analoog-digitaalomzetting. Het signaal wordt bemonsterd met een samplefrekwentie f_s die, zoals gezegd, minstens het dubbele moet bedragen van de hoogste frekwentie in het spraaksignaal. De keuze van f_s ligt in principe bij de gebruiker, waarbij de bovengrens wordt bepaald door de tijd die nodig is om de samples weg te schrijven naar het schijvengeheugen van de computer. Bij de IPO-P857 computer is f_s ten hoogste 12 kHz; bij de IPO-VAX is het tienvoudige haalbaar. Als standaardwaarde gebruiken we f_s=10 kHz; een andere veel gebruikte (normale) waarde is 8 kHz.

De analoog-digitaalomzetter heeft een (standaard-)precisie van 12 bits. Ieder sample kan dus een gehele getalwaarde hebben tussen -2048 en +2047. Bij volle uitsturing bedraagt de verhouding tussen signaal en kwantiseringsruis dus ongeveer 66 dB. De 12 bits samples worden opgeslagen in 16 bits integer woorden en bij een f_s van 10 kHz vereist iedere seconde spraak dus 10000 integer woorden, waarvan 4 bits per

woord ongebruikt zijn.

2.3.2. De eigenlijke analyse

Voor de analyse wordt een spraaksegment van standaardlengte 25 ms, dat bij een f_s van 10 kHz dus 250 samples bevat, van de schijf gelezen. Op de grootte van dit analysevenster komen we nog terug in hoofdstuk 4 en 5.

Vooraf bepalen we of er voldoende signaal in het analysevenster aanwezig is. Pauzes en andere stukken waar het signaal (vrijwel) nul is worden niet verder geanalyseerd. Alleen als de totale energie R_0 van het ingangssignaal groter is dan een bepaalde drempelwaarde, volgt de eigenlijke analyse. Van de bronparameters wordt dan eerst de stem/stemloos parameter bepaald, dus vóórdat enige bewerking op het signaal is uitgevoerd. Daartoe wordt van het segment binnen het analysevenster de verhouding R_1/R_0 volgens (2.2.1) berekend. Als die hoger is dan een van beide grenswaarden zoals uiteengezet is in par. 2.2.1, wordt het segment als stemhebbend geklassificeerd, anders stemloos.

Daarna worden alle samples s_n binnen het analysevenster vermenigvuldigd met een faktor:

(2.3.1) $\quad m_n = .54 - .46 \cos(2\pi n/N)$,

waarin N het totale aantal samples in het analysevenster is (standaard is N = 250). Dit vermenigvuldigen met een Hammingwindow wordt gedaan om artefakten in de vorm van spectrale zijlobben ('transiënten') zoveel mogelijk te onderdrukken. Als er geen, dwz een rechthoekige windowfunktie zou worden toegepast kunnen in delen waar het ingangsspectrum een lage amplitude heeft de zijlobben overheersen en zou in die gebieden niet de spectrale omhullende van het spraaksignaal maar die van de transiënten worden geneutraliseerd. De vorm van de toe te passen windowfunktie is niet kritisch zolang de overgangen aan de uiteinden maar niet al te abrupt zijn. In de praktijk voldoet het hier toegepaste Hammingwindow, met zijlobben lager dan 43 dB, uitstekend.

De volgende bewerking die dan wordt toegepast is de zgn pre-emfase, waarbij ieder sample s_n wordt vervangen door s_n' met:

(2.3.2) $\quad s_n' = s_n + u\, s_{n-1}, \qquad$ met $u = -.9$.

Dit is een vast pre-emfase filter:

(2.3.3) $\quad P(z) = 1 + u\, z^{-1}$, $\qquad$ met $u = -.9$,

dat zowel bij stemhebbende als bij stemloze spraaksegmenten wordt toegepast. Het verzwakt de lage frekwenties t.o.v. de hoge en de overdrachtsfunktie heeft bij benadering een helling van +6 dB/oct. Deze pre-emfase wordt toegepast om de lange-termijn gemiddelde helling van -6 dB/oct te kompenseren. Op de gevolgen hiervan komen we nog terug in hoofdstuk 4.

2.3.3. Filterberekening

Vervolgens worden van dit aldus voorbewerkte signaal de eerste M+1 autocorrelaties berekend volgens:

$$(2.1.4a) \qquad R_i = \sum_{n=1}^{N} s_n \, s_{n-i} \, , \qquad i = 0 \ldots M$$

met N het aantal samples in het analysevenster (standaard 250 stuks) en M het aantal te berekenen filtercoëfficiënten dat in principe door de gebruiker vrij kan worden gekozen. Standaard is M = 10 en op de keuze hiervan komen we nog terug in hoofdstuk 4.

Met deze autocorrelaties wordt vervolgens het stelsel vergelijkingen (2.1.5) recursief opgelost. Het resultaat is dan een set filtercoëfficiënten, de a-parameters, die de impulsresponsie van het analysefilter, en daarmee dus ook de overdrachtsfunktie van dat filter, voor het betreffende spraaksegment volledig bepalen. In een aantal situaties kan met deze a-parameters worden volstaan. Meestal is echter een beschrijving van het analysefilter gewenst als produkt van 2^e orde filters. De volgende stap in de analyse is dan ook het afsplitsen van kwadratische termen van het a-polynoom, hetgeen de pq-parameters als de coëfficiënten van de 2^e-orde filters oplevert. Indien gewenst worden deze dan nog gesorteerd en omgezet in coëfficiënten van louter resonerende deelfilters, zoals in par. 2.5 uiteengezet zal worden. Daarmee is dan de analyse van het betreffende spraaksegment, voor wat betreft de filterparameters, voltooid.

Van de bronparameters hebben we de stem/stemloosbeslissing al bepaald. De amplitudeversterkingsfaktor G wordt als 'bijprodukt' van de analyse verkregen uit de energie E_m van het restsignaal volgens (2.1.6). Tenslotte wordt, voor zover althans het spraaksegment als stemhebbend is geklassificeerd, de periode van de grondtoon bepaald. Daartoe wordt een analysevenster gebruikt dat langer is dan 25 ms zoals gebruikt bij de filtercoëfficiënten. Dit hangt samen met het feit dat we ook bij lage grondtoonfrekwenties van bv 50 Hz nog minstens 2 perioden van de grondtoon van het spraaksignaal in het venster willen hebben, om die periode voldoende nauwkeurig te kunnen bepalen

(Duifhuis et al, 1982). Bij de toonhoogtemeting wordt dan ook gewerkt met een standaard vensterlengte van 40 ms, dus 400 samples bij 10 kHz samplefrekwentie.

Hiermee is dan voor het betreffende spraaksegment de analyse kompleet en de verkregen set parameters, het frame, wordt in een file op het schijvengeheugen opgeslagen.

Dan wordt het volgend stukje spraak in het analysevenster geplaatst en de gehele cyclus herhaald. Daarbij kan de stapgrootte ofwel de frameperiode, dat is de tijdsduur waarover het analysevenster wordt 'opgeschoven' in het spraaksignaal, in principe door de gebruiker vrij worden gekozen. In de praktijk wordt een standaard frameperiode van 10 ms (100 samples bij 10 kHz) gekozen, waarbij dus een overlapping tussen twee opeenvolgende spraaksegmenten optreedt van 15 ms. Ook op de keuze van deze frameperiode komen we in hoofdstuk 4 en 5 nog terug.

Aldus wordt de gehele spraakuiting in vaste stappen doorlopen waarbij de opeenvolgende frames, met de bij elkaar horende parameters, na elkaar in een file van het schijvengeheugen worden weggeschreven. Van het resultaat van zo'n analyse zullen we in de volgende paragraaf een paar voorbeelden bespreken.

2.4. RESULTATEN: RESOGRAMMEN

Twee voorbeelden van de resultaten van een (standaard)analyse zijn grafisch weergegeven in fig. 2.7, voor de zinnen "ieder half uur komt hier een bus langs" en "de bal vloog over de schutting", uitgesproken door resp. een mannen- en een vrouwenstem. De inhoud van de opeenvolgende frames is hier als funktie van de tijd weergegeven, 100 frames voor 1 seconde spraak. In de bovenste helft van beide plaatjes staan de bronparameters, met van boven naar beneden: amplitudeversterkingsfaktor G, stemloos-markering UV en grondtoonfrekwentie F_0. In de onderste helft zijn de FQ-parameters van de afzonderlijke resonanties van het analysefilter uitgezet in wat we verder het 'resogram' zullen noemen.

In de fonetiek is het gebruikelijk om energiespectra weer te geven als funktie van de tijd door middel van een spectrogram of 'sonagram'. Daarin is in de tijd (horizontaal) uitgezet hoe de energie verloopt als funktie van de frekwentie (verticaal), waarbij het grijsnivo een maat is voor de energie binnen een bepaald frekwentiegebied (hoe donkerder hoe meer energie).

Ons resogram is enigszins te vergelijken met een (vereenvoudigd) spectrogram, maar dan zonder grijstinten. In principe zou in het reso-

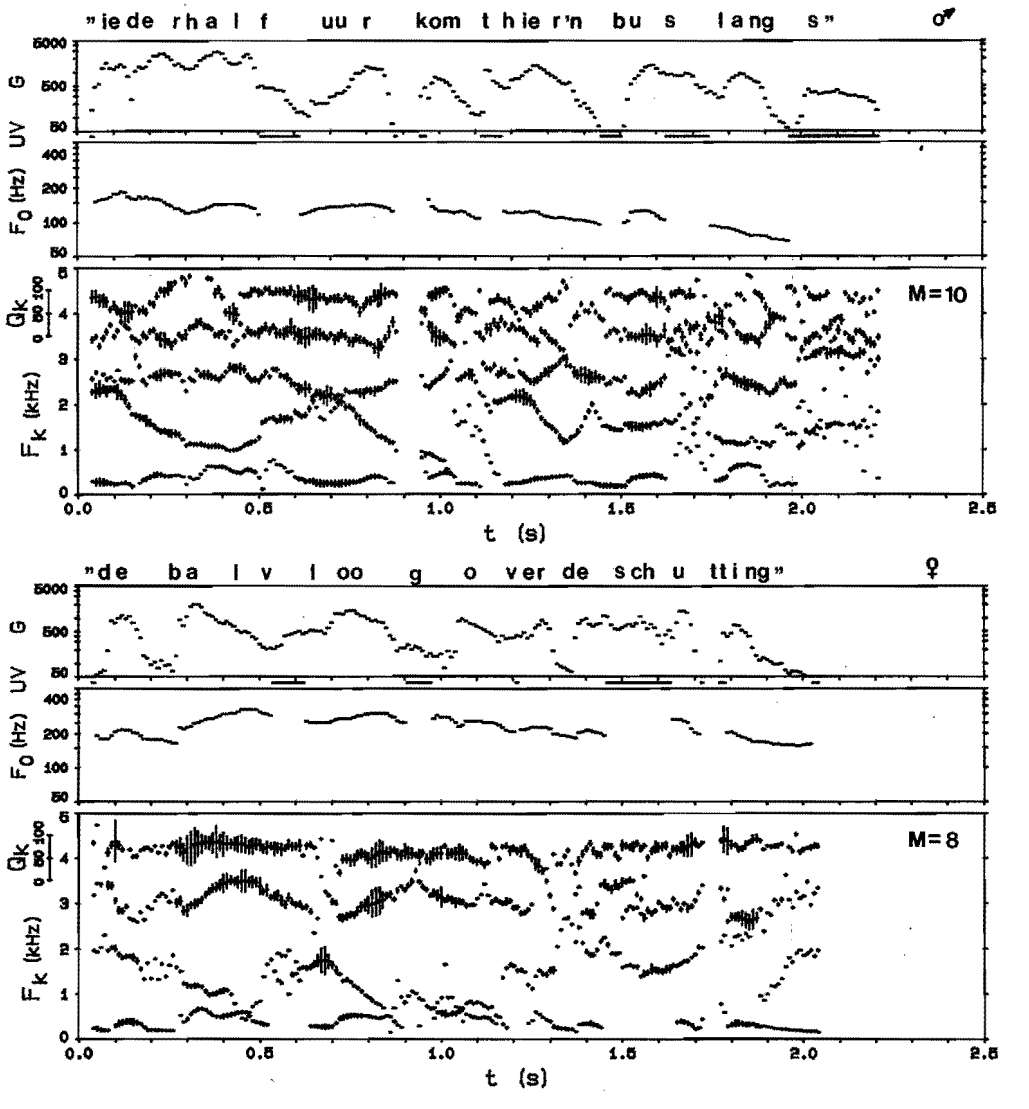

Fig.2.7. Analyseresultaten voor een mannenstem (boven), geanalyseerd met M = 10 filtercoëfficiënten en een vrouwenstem (onder) met M = 8. Van boven naar beneden: amplitudeversterkingsfaktor G, stemloosindicatie UV (zwart = stemloos) en grondtoonfrekwentie F_0 van het bronsignaal. Daaronder het resogram met de afstemfrekwenties F_k met daaromheen de kwaliteitsfaktoren Q_k. De totale lengte van de verticale streepjes geeft Q_k weer, volgens de aparte schaal links.

gram, naast de afstemfrekwentie F van de afzonderlijke deelfilters, de bandbreedte B als funktie van de tijd kunnen worden weergegeven. Het nadeel daarvan is echter dat dan breedbandige resonanties veel meer in het oog zouden springen dan smalbandige, die perceptief veel relevanter zijn. Daarom is in het resogram in plaats van de bandbreedte B de kwaliteitsfaktor Q = F/B uitgezet, gecentreerd rond de afstemfrekwentie F. Lange vertikale strepen in het resogram betekenen dus een scherpe, selectieve (anti)resonantie met hoge Q. Deze zijn te associëren met hoge, smalle pieken in de spectrale omhullende van het ingangssignaal en met de donkere gebieden in het spectrogram. De lengte van de Q-strepen heeft echter geen direkte betekenis op de frekwentieas; daarvoor geldt de (apart weergegeven en dimensieloze) Q-schaal. Secties waarvoor geen FQ-paar is gedefinieerd, omdat de bijbehorende nulpunten reëel zijn, zijn vooralsnog weggelaten. We komen daarop terug in par. 2.5.

In fig. 2.7 zijn bij stemhebbende stukken en vooral bij klinkers, duidelijk samenhangende trajecten in de verschillende afstemfrekwenties te onderscheiden. Stemloze stukken zijn doorgaans minder duidelijk gestruktureerd. We zien ook hoe bij de overgrote meerderheid van de frames het aantal antiresonanties precies de helft is van het aantal filtercoëfficiënten. Voor die frames hebben de afzonderlijke 2^e-orde secties kennelijk alle toegevoegd complexe nulpunten. Reële nulpunten komen vooral voor bij stemloze en nasale fragmenten, waar we nogal wat frames zien met minder dan M/2 antiresonanties.

Als we de resultaten van de mannen- en vrouwenspraak onderling vergelijken zien we duidelijk hoe het onderste resogram in fig. 2.7 minder dicht is gevuld met dikke contouren, dus resonanties met een hoge Q. Dat hangt samen met het feit dat in het spectrum tot 5 kHz bij vrouwenspraak doorgaans een formant minder is te onderscheiden dan bij mannenspraak. De analyse is ook met M = 8 i.p.v. 10 uitgevoerd. We komen daarop nog uitvoerig terug in hoofdstuk 4 en 5.

Wat de bronparameters betreft zien we dat de grondtoonfrekwentie veel geleidelijker verloopt dan de andere parameters en dat de gemiddelde ligging bij de vrouwenstem hoger is dan bij de mannenstem.

In de hier gepresenteerde resogrammen zijn de deelfilters waarvan de nulpunten niet toegevoegd complex zijn niet weergegeven. Voor onder meer zuinige codering van het spraaksignaal is het echter gewenst om de analyseresultaten in een zodanige vorm te gieten dat alle M/2 secties toegevoegd complexe nulpunten hebben, zodat voor ieder frame altijd precies M/2 FQ-paren terug te vinden zijn in het resogram. Hoe we dat doen wordt in de volgende paragraaf uiteengezet.

2.5. TRANSFORMATIE NAAR LOUTER RESONERENDE DEELFILTERS

2.5.1. Probleemstelling

In de vorige paragraaf hebben we de analyseresultaten weergegeven in de vorm van resogrammen, waarin het M^e orde analysefilter door ten hoogste M/2 afstemfrekwenties F en kwaliteitsfaktoren Q van afzonderlijke deelfilters is gekarakteriseerd. We hebben echter ook gezien dat deze FQ-paren alleen gedefinieerd zijn voor filters met toegevoegd complexe nulpunten in de overdrachtsfunktie. De niet-resonanties met reële nulpunten in de overdrachtsfunktie zijn dus nog niet in het resogram gerepresenteerd. Zowel voor het experimentele spraakonderzoek als voor toepassingen waarbij de spraak zuinig gecodeerd wordt is het nodig om ieder frame met uitsluitend resonanties te beschrijven. De eerste vraag waarop we in deze paragraaf zullen ingaan is dan ook: hoe kunnen we pq-paren die in de overdrachtsfunktie van het analysefilter reele nulpuntenparen representeren, omzetten in paren met toegevoegd complexe nulpunten en daarbij zo weinig mogelijk veranderen in het energiespectrum van zo'n deelfilter?

Behalve het negeren van de secties met reële nulpunten is er in de resogrammen zoals die tot nu toe zijn gepresenteerd nog een tweede moeilijkheid verborgen. Die is dat de parameters van de opeenvolgende frames niet op elkaar hoeven aan te sluiten. Ze zijn in principe namelijk nog niet geordend langs de frekwentieas. Dat komt omdat bij het afsplitsen uit het A-polynoom de volgorde waarmee de pq-paren ter beschikking komen willekeurig kan zijn. Weliswaar zal bij een goede beginschatting voor de iteraties veelal een pq-paar beschikbaar komen dat dicht bij die schatting (b.v. van het vorige frame) ligt, maar noodzakelijk is dat niet. Het hangt van min of meer toevallige faktoren af welke kwadratische term bij het iteratieproces het eerst tot convergentie leidt en wordt afgesplitst. Dat betekent ook dat de afstemfrekwenties van de antiresonanties die uit deze ongeordende pq-paren zijn berekend evenmin naar frekwentie zijn geordend. Aan de resogrammen is dat niet te zien zolang de parameterwaarden van de opeenvolgende frames niet met elkaar worden doorverbonden. Maar wanneer we tussen pq-paren van opeenvolgende frames willen interpoleren, hetgeen met name voor zuinige codering van het spraaksignaal nodig kan zijn, moeten de parameterwaarden zo goed mogelijk op elkaar aansluiten en is dus een of andere vorm van ordening noodzakelijk.

De tweede vraag die we in deze paragraaf behandelen is dan ook: hoe kunnen we de pq-parameters ordenen langs een frekwentieschaal. Simpelweg sorteren naar toenemende p-waarde is daarvoor niet toereikend, aangezien de afstemfrekwentie F zowel van p als van q afhangt volgens (2.1.40).

Dit is geïllustreerd in fig. 2.8, waarin de mogeljke ligging van de pq-paren is geschetst. Een paar waarvan de bandbreedte afneemt en de afstemfrekwentie F konstant blijft, verplaatst zich in het pq-vlak langs een parabool, b.v. van b' naar b. Zo'n breedbandige resonantie b, waarvan de afstemfrekwentie F lager is dan die van een smalbandige resonantie a kan dus best een p-coëfficiënt hebben die groter is dan die van de smalbandige. Het is dus nodig eerst om te rekenen naar afstemfrekwentie en bandbreedte en pas dan kunnen we sorteren naar toenemende F. Maar dat werpt dan de vraag op wat te doen met pq-paren waarvoor geen bijbehorend FQ-paar is gedefinieerd.

2.5.2. Transformatie van pq- naar cr-parameters

Om beide problemen: a) hoe te ordenen en b) wat te doen met 'niet-resonanties' aan te pakken worden de pq-paren getransformeerd naar zgn cr-parameters, een tussenstation op weg naar FB- of FQ-paren (Willems en Vogten, 1979).
Deze cr-parameters worden gedefinieerd door:

(2.5.1) $c = p / \sqrt{|q|}$, $q \neq 0$

en

(2.5.2) $r = \mathrm{sgn}(q) \sqrt{|q|}$,

zodat ook:

(2.5.3) $p = \mathrm{sgn}(r)\, rc$

en

(2.5.4) $q = \mathrm{sgn}(r)\, r^2$.

De paraboolschijf in het pq-vlak in fig. 2.8 die wordt begrensd door $p^2 = 4q$ en de rechte $q = 1$ gaat dan over in de rechthoek in het cr-vlak, begrensd door $|c| = 2$, $r = 1$ en $r = 0$. 'Niet-resonerende' pq-paren met een positieve q waarde, die op of onder de parabool liggen, komen in het cr-vlak dus op of naast de rechthoek terecht.

Paren met negatieve q liggen in het negatieve deel van het cr-vlak. Resonerende pq-paren worden getransformeerd naar resonerende cr-paren, die in het cr-vlak binnen de rechthoek liggen. Voor deze paren, corresponderend met toegevoegd complexe nulpunten in de overdrachtsfunktie, is het verband met de afstemfrekwentie F en bandbreedte B gegeven door:

(2.5.5) $F = (1/2\pi T) \arccos(-c/2)$, $|c| < 2$

en

(2.5.6) $B = (-1/\pi T) \ln(r)$, $r > 0$,

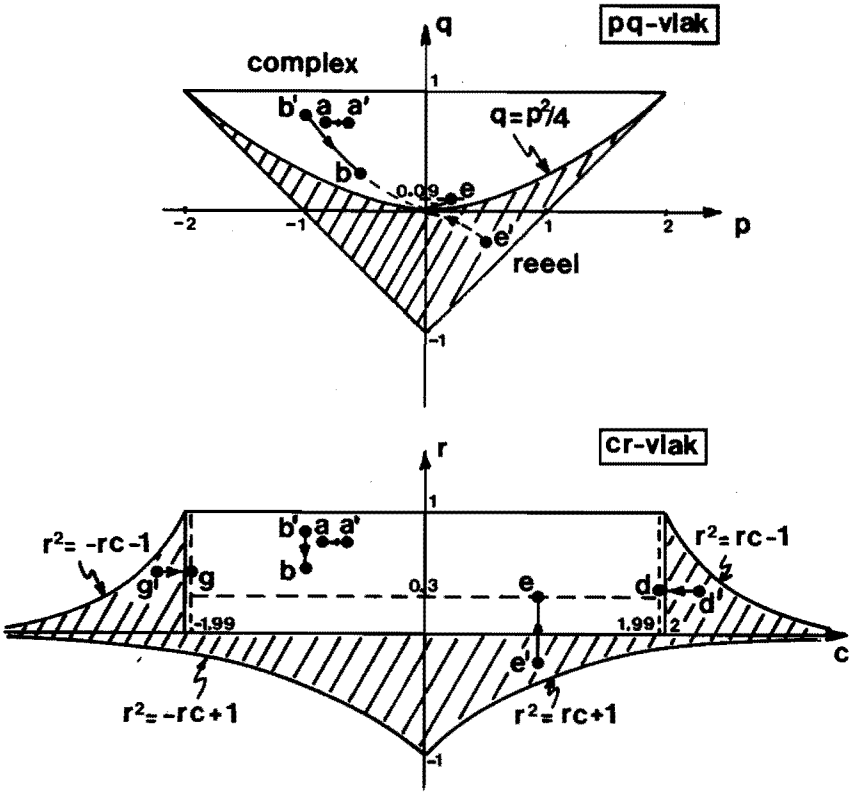

Fig.2.8. Verband tussen pq- en cr-parameters.

---

zodat ook:

(2.5.7) $c = -2 \cos (2\pi FT)$

en

(2.5.8) $r = \exp (-\pi BT)$,

waarin T weer de bemonsteringsperiode $1/f_s$ voorstelt.

Deze transformatie heeft dus tot gevolg dat de afstemfrekwentie alleen van c afhangt, niet van r. Bij konstante F leidt een toenemende bandbreedte in het cr-vlak bijvoorbeeld tot een verplaatsing van b' naar b, loodrecht op de c-as.

Het sorteerprobleem kunnen we na deze transformatie oplossen door alle M/2 cr-paren, dus ook de niet-resonerende, te ordenen naar toenemende c-waarde. Dat levert een sortering naar toenemende frekwenties, zij het dat bij 'niet-resonerende' cr-paren dit een frekwentie kan

zijn die negatief is of groter dan de halve samplefrekwentie f_s. Afstemfrekwentie F en bandbreedte B zijn volgens (2.5.5 en 6) alleen gedefinieerd voor cr-paren die binnen de rechthoek c = ±2, r = 0 en r = 1 liggen. Voor c-waarden kleiner dan -2 zouden we in principe negatieve afstemfrekwenties kunnen toelaten en voor c groter dan 2 afstemfrekwenties groter dan $f_s/2$. Zo ook zouden we voor negatieve r de definitie van bandbreedte kunnen uitbreiden door ook negatieve waarden voor B toe te laten. Dat doen we hier niet, omdat daarmee dan toch nog geen eenvoudige representatie in resogrammen mogelijk is, met een hanteerbare indeling naar opeenvolgende resonanties.

Om ervoor te zorgen dat alle cr-paren corresponderen met antiresonanties worden de naar c geordende paren begrensd, zodanig dat ze binnen de rechthoek in het cr-vlak vallen (Willems en Vogten, 1979), dus:

(2.5.9) $|c| \leq 1.99$

en

(2.5.10) $r \geq 0.3$.

Niet-resonerende paren worden in het cr-vlak in fig. 2.8 dus verschoven tot net binnen de rechthoek, waarbij negatieve c-waarden worden begrensd tot -1.99 en positieve c-waarden tot +1.99. In het cr-vlak verschuiven bv. de punten g' en d' dan naar g en d, waarmee de afstemfrekwentie net positief wordt cq. net onder de halve samplefrekwentie komt te liggen. Tevens worden de r-waarden, om praktische redenen, begrensd tot minimaal 0.3, zodat de grootst mogelijke bandbreedte ruim onder de halve samplefrekwentie komt te liggen. In het cr-vlak verschuift bv. het punt e' dus naar e.

Daarmee zijn de cr-paren in de overdrachtsfunktie veranderd in resonerende paren met toegevoegd complexe nulpunten. Aldus is voor alle M/2-paren een afstemfrekwentie F en een (grote) bandbreedte B gedefinieerd volgens (2.5.5 en 6) en zijn voor ieder frame de parameters van het filter geordend naar toenemende afstemfrekwentie en weer te geven in het resogram.

In fig. 2.9 is een voorbeeld gegeven van resogrammen voor en na deze sorteer- en begrenzingsprocedure, voor het zinsfragment "komt hier 'n bus langs". In het oorspronkelijke resogram (bovenste plaatje) ontbreken nogal wat antiresonanties. Alle frames die een of meer pq-paren bevatten met reële nulpunten zijn gemerkt. Het onderste resogram geeft het resultaat weer nadat de hele procedure van sorteren en begrenzen is doorlopen. Alle frames bestaan nu uit precies 5 resonanties. We zien dan ook hoe in de stemloze s van "bus" en "langs" en ook in de nasalen m van "komt", n van "'n bus" en ng van "langs" er in het onderste resogram nogal wat punten bijgekomen zijn.

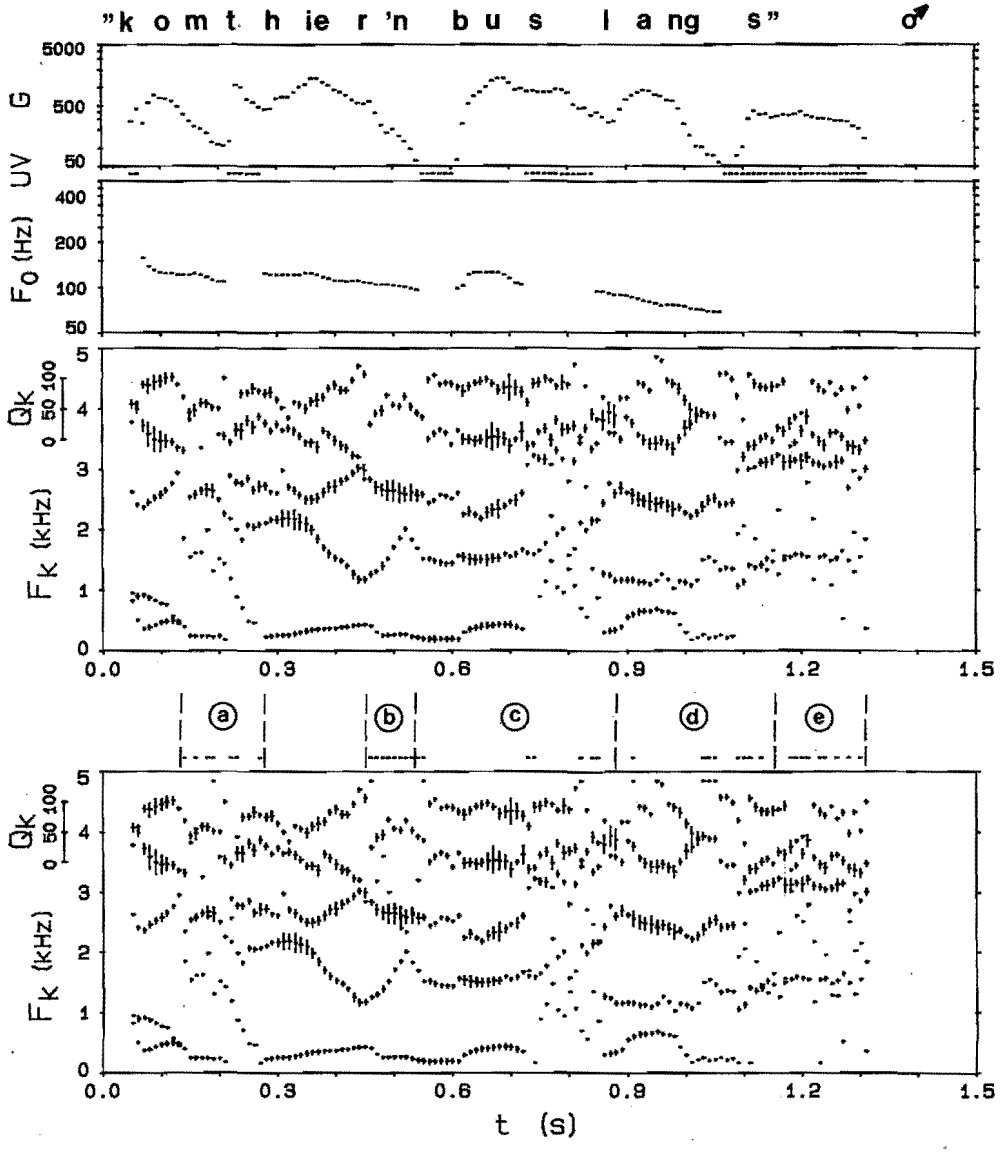

Fig.2.9. Bovenste 3 plaatjes: analyseresultaten voor M = 10 van het fragment "komt hier 'n bus langs" uit fig. 2.7 boven. Onderste plaatje: resogram van hetzelfde fragment nadat alle deelfilters resonerend zijn gemaakt. Alle frames met oorspronkelijk minder dan 5 resonerende deelfilters (vooral voorkomend bij de nasalen m, n en ng en in de stemloze stukken) zijn gemarkeerd bij (a) t/m (e) en hun spectra zijn in fig. 2.10a t/m e afzonderlijk (gestippeld) weergegeven.

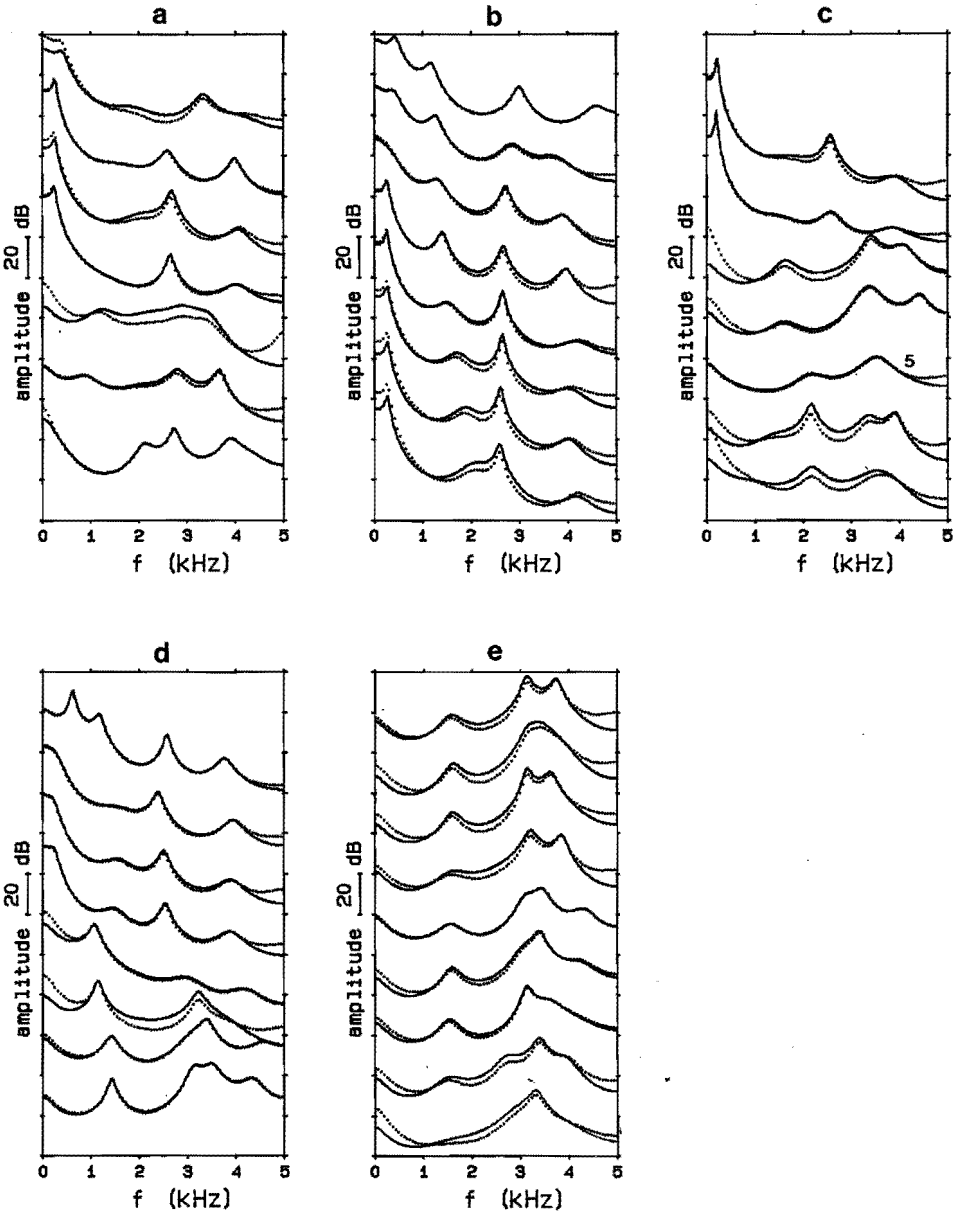

Fig.2.10. Stippellijnen: oorspronkelijke (geïnverteerde) energiespectra van alle frames uit fig. 2.9, die minder dan 5 resonerende deelfilters bevatten. Getrokken lijnen: spectra van dezelfde frames nadat alle deelfilters resonerend zijn gemaakt via de sorteer- en begrenzingsprocedure.

2.5.3. Gevolgen en conclusie

Dat roept dan onmiddellijk de vraag op welke gevolgen deze procedure heeft op het energiespectrum van het aldus veranderde analysefilter. Hoe gering die zijn zien we in fig. 2.10, waarin voor alle in fig. 2.9 gemerkte frames (met oorspronkelijk een of meer niet-resonerende deelfilters) het (geïnverteerde) analysefilter is weergegeven. De gestippelde curves zijn steeds de oorspronkelijke energiespectra, berekend via de Fouriertransformatie van de impuls-responsie van het analysefilter. De getrokken curves zijn de spectra verkregen na afsplitsen van de pq-paren volgens (2.1.20), transformatie van cr-paren volgens (2.5.1 en 2), sorteren en begrenzen volgens (2.5.7 t/m 10), terugtransformeren volgens (2.5.3 en 4) en uitvermenigvuldigen volgens (2.1.20) tot het A-polynoom.

We zien dat de eventuele veranderingen die deze procedure veroorzaakt in het energiespectrum meestal gering zijn. Soms zien we een wat grotere afval aan de hoge kant van het spectrum, soms aan de lage kant. In het algemeen is echter, door het feit dat de nieuwe resonanties een grote bandbreedte krijgen, het effekt van deze verandering in de overdrachtsfunktie van de deelfilters op het resulterende analysefilter klein. Zelfs voor frame nr 5 in fig. 2.10c, waar twee van de vijf resonanties in het oorspronkelijke (gestippelde) spectrum ontbreken, is het verschil tussen beide curves zeer klein. De grote bandbreedte maakt dat ook in het spectrum met louter resonanties (getrokken curve) maar twee van de vijf aanwezige resonanties als duidelijke pieken in het energiespectrum terug te vinden zijn.

Uit het bovenstaande kunnen we concluderen dat de hier gepresenteerde ordenings- en begrenzingsprocedure waarmee alle nulpunten in de overdrachtsfunktie van het analysefilter toegevoegd complex en dus alle deelfilters resonerend worden gemaakt, in het algemeen weinig invloed heeft op het resulterende energiespectrum.

Dat deze procedure ook geen hoorbaar effekt heeft op de uiteindelijk geresynthetiseerde spraak zal in hoofdstuk 5 blijken. Eerst zullen we in het volgende hoofdstuk uiteenzetten hoe we met de verkregen analyseresultaten de spraak weer reconstrueren: de resynthese.

3 RESYNTHESE VAN HET SPRAAKSIGNAAL

In dit hoofdstuk komt aan de orde hoe we spraaksignalen uit de berekende analyseresultaten kunnen resynthetiseren. Basis voor deze resynthese is het bron-filtermodel voor de fysica van de menselijke spraakproduktie, zoals in hoofdstuk 1 is besproken.
In het vorige hoofdstuk hebben we behandeld hoe de parameters van het analysefilter zó worden berekend dat voor het betreffende spraaksegment in het analysevenster het uitgangssignaal van het filter een zo vlak mogelijk energiespectrum heeft. Gelet op het feit dat in het produktiemodel beide bronsignalen een vlak, wit energiespectrum hebben, kan dan het oorspronkelijke spraaksignaal benaderd worden door het berekende analysefilter te inverteren en het aldus verkregen synthesefilter te exciteren met een periodieke impuls of met witte ruis.

Het eerste deel van dit hoofdstuk beschrijft deze resynthese en daarin wordt tevens aangegeven hoe de amplitude(versterkingsfaktor) kan worden verkregen.

In het tweede deel wordt de praktische uitvoeringsvorm van de resynthese besproken, en we maken daarin onderscheid tussen twee werkwijzen. Bij de eerste (asynchrone) methode gebeurt het vernieuwen van de filterparameters op vaste tijdstippen, gegeven door de frameperiode die bij de analyse is toegepast. De periodiekbron genereert geheel los daarvan impulsen met tussenpozen die bepaald worden door de gemeten periodeduur van de grondtoon (toonhoogte), zodat het filter dus asynchroon met de grondtoonperiode wordt bijgesteld. In de tweede (synchrone) methode worden de filterparameters steeds aan het begin van een nieuwe grondtoonperiode bijgesteld, dus synchroon met die periode. Bij de laatste methode worden tevens de filterparameters van twee opeenvolgende frames geinterpoleerd, waardoor in sommige gevallen de geresynthetiseerde spraak iets beter kan klinken. Hoewel deze synchrone methode theoretisch de voorkeur heeft kost ze wat meer rekentijd dan de asynchrone, maar beide staan ter beschikking van de gebruiker.

3.1. BEPALING SYNTHESEFILTER EN AMPLITUDE

Bij de analyse hebben we de coëfficiënten van een Me orde filter A berekend door de uitgangsenergie te minimaliseren. Dit analysefilter had als overdrachtsfunktie in het z-domein:

$$A(z) = \sum_{k=1}^{M} a_k \, z^{-k} \quad , \tag{2.1.8}$$

en het uitgangssignaal e_n werd in het tijddomein gegeven door:

$$e_n = s_n + \sum_{k=1}^{M} a_k \, s_{n-k} \quad , \tag{2.1.1}$$

als som van ingangssignaal op tijdstip n en de gewogen M voorafgaande ingangssamples.

De coëfficiënten $/a_k/$ zijn zo bepaald dat de spectrale omhullende van het signaal aan de uitgang van het filter, voor gegeven M, zo vlak mogelijk is geworden. Dat betekent dat deze overdrachtsfunktie A(z) de geïnverteerde omhullende van het ingangssignaal zo goed mogelijk benadert. Dus beschrijft 1/A(z) de spectrale omhullende van het ingangssignaal zo goed mogelijk. Als synthesefilter definiëren we dan ook:

$$H(z) = 1 \; / \; A(z) \quad . \tag{3.1.3}$$

Het ingangssignaal voor dit synthesefilter is een bronsignaal u_n met als z-getransformeerde U(z) en het uitgangssignaal is een signaal s_n', met z-getransformeerde S'(z), dat het spraaksignaal aan de ingang van het analysefilter zo goed mogelijk benadert (fig. 3.1).
In het z-domein geldt dus:

$$S'(z) = U(z) \; H(z)$$

of

$$S'(z) = U(z)/A(z) = U(z) \; /(1 + \sum_{k=1}^{M} a_k \, z^{-k})$$

Terugtransformatie naar het tijddomein levert:

$$s_n' + \sum_{k=1}^{M} a_k \, s_{n-k}' = u_n \quad ,$$

of:

$$s_n' = u_n - \sum_{k=1}^{M} a_k \, s_{n-k}' \quad . \tag{3.1.4}$$

Het uitgangssignaal wordt dus gegeven door het verschil van ingangssignaal op tijdstip n en een lineaire kombinatie van M daaraan voorafgaande uitgangssamples, fig. 3.2.

Wat is nu het ingangssignaal u_n voor dit synthesefilter? Vergelij-

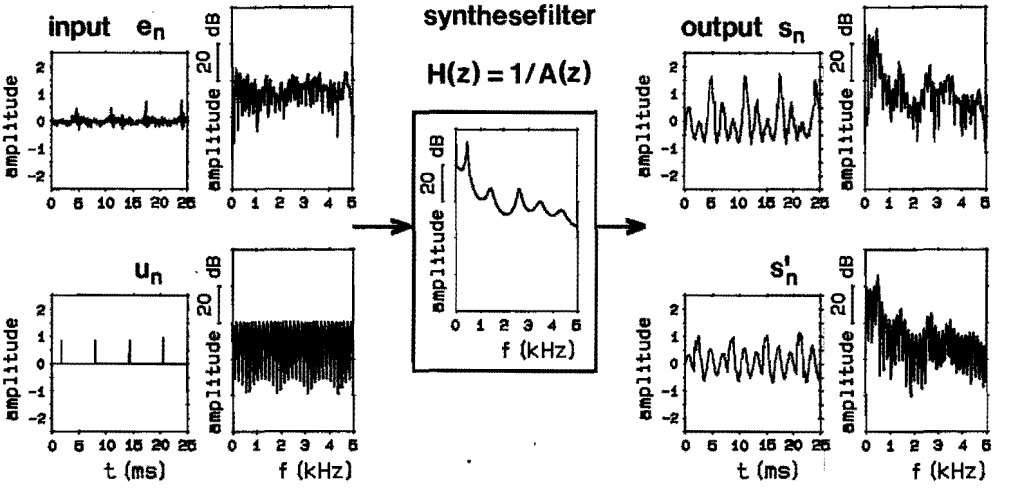

Fig.3.1a. Energiespectrum (midden) van het synthesefilter H(z), verkregen door het analysefilter A(z) uit fig. 2.2 (periodiek ingangssignaal) te inverteren. Het restsignaal e_n (links boven) als input voor dit filter levert het oorspronkelijke spraaksignaal s_n op (rechts boven). Bij de resynthese wordt echter niet e_n maar een impulsreeks u_n (links onder) als ingangssignaal toegepast. Dat levert een synthetisch spraaksignaal s_n' (rechts onder).

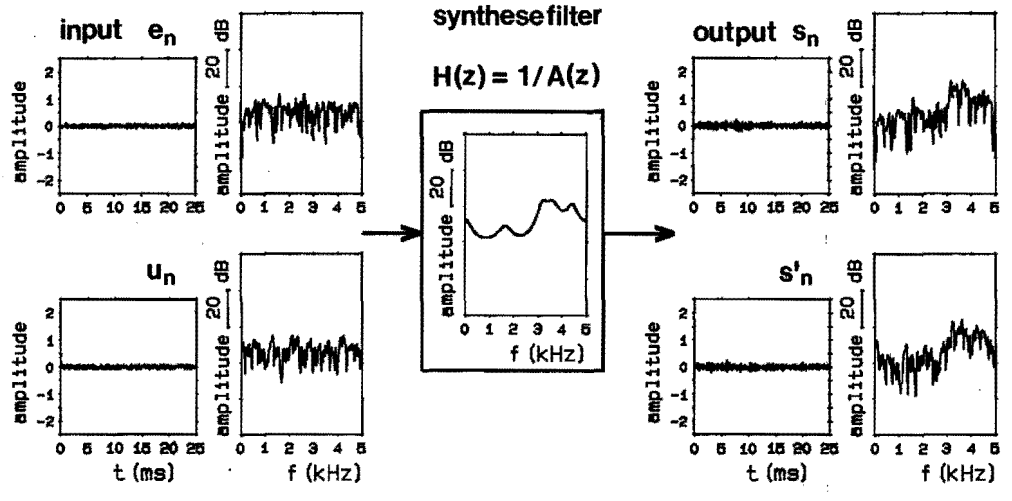

Fig.3.1b. Als Fig. 3.1a, maar nu voor het filter uit fig. 2.3, berekend voor een ruisig (stemloos) ingangssignaal. Bij de resynthese wordt niet het restsignaal e_n (links boven), maar witte ruis (links onder) als ingangssignaal voor het synthesefilter toegepast.

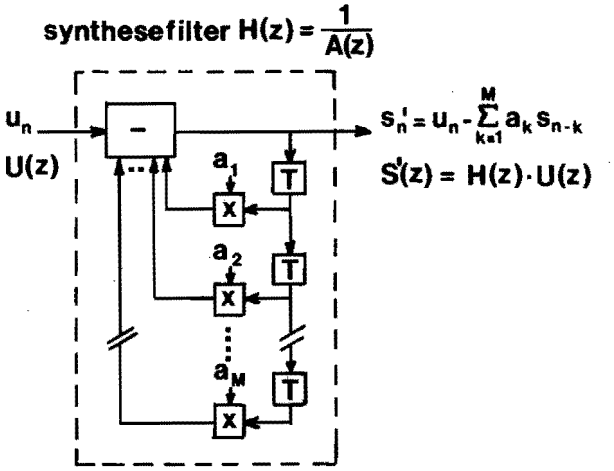

Fig.3.2. Digitaal Me orde synthesefilter, verkregen door inverteren van het analysefilter. Het uitgangssample s_n' op tijdstip nT wordt gegeven door het verschil tussen het ingangssample u_n op datzelfde tijdstip en M gewogen daaraan voorafgaande uitgangssamples. De weegfaktoren /a_k/ zijn de a-parameters van het berekende analysefilter.

---

lijking van (3.1.4) met (2.1.1) leert dat wanneer we als input voor het synthesefilter het restsignaal e_n, dus het uitgangssignaal van het analysefilter zouden nemen, we daarmee als output s_n' van het synthesefilter weer het oorspronkelijke spraaksignaal s_n verkrijgen. Dat is in fig. 3.1 is geïllustreerd en moet natuurlijk ook zo zijn; het produkt van A(z) en H(z) is immers per definitie 1. Voor de praktijk van het spraakonderzoek is dit dan ook geen zinvolle werkwijze. We kunnen het restsignaal e_n moeilijk opvatten als resultaat van een analyse "in termen van een beperkt aantal perceptief relevante parameters", zoals we in hoofsdtuk 0 hebben geeist. Ook bevat het restsignaal e_n bijna evenveel informatie als het oorspronkelijke spraaksignaal s_n en samen met die van het analysefilter zou dat dus geen enkele reduktie inhouden van de datastroom.

Wanneer we echter uitgaan van het geïdealiseerde geval dat het restsignaal e_n een echt wit spectrum heeft, dan kan het restsignaal bestaan uit impulsen of ongecorreleerde ruis. Welnu, we passen hier deze geïdealiseerde restsignalen toe als ingangssignaal u_n. Conform het produktiemodel van hoofdstuk 1 is het bronsignaal dus ofwel een periodieke impuls met herhalingsfrekwentie F_o ofwel witte ruis, afhankelijk van de stem/stemloosparameter zoals die bij de analyse is bepaald.

Het voor de resynthese 'ideale' brongeluid, het restsignaal e_n,

wordt dus vervangen door een sterk vereenvoudigde versie daarvan. We merken op dat hierdoor niet alleen de nog resterende amplitude-informatie over het oorspronkelijke spraaksignaal verloren gaat, maar ook alle faseinformatie die daarover in dat restsignaal aanwezig is. Het analysefilter zelf bevat geen faseinformatie over de oorspronkelijke spraak; het is immers berekend op grond van alleen het energiespectrum.

Rest ons nog de specificatie van de versterkingsfaktor G waarmee in het model de energie van het spraaksegment kan worden geregeld. Deze wordt bepaald door de eis dat de energie van het signaal na resynthese gelijk moet zijn aan die van het oorspronkelijke ingangssignaal van het analysefilter.

Laten we uitgaan van een ingangssignaal u_n voor het (causale) synthesefilter dat uit één enkele eenheidsimpuls bestaat en vermenigvuldigen we dit met de (nog onbekende) versterkingsfaktor G. Dan wordt de (impuls)responsie van het synthesefilter gegeven door (3.1.4), dus:

$$s_n' = G\, d_n - \sum_{k=1}^{M} a_k\, s_{n-k}' \quad \text{met } d_n=1 \text{ als } n=0 \text{ en } d_n=0 \text{ elders.}$$

Om de totale energie van dit signaal te bepalen kunnen we met s_{n-i}' vermenigvuldigen en sommeren over alle n (Makhoul,1975). Dat levert:

$$\sum_n s_n' s_{n-i}' = G \sum_n d_n\, s_{n-i}' - \sum_{k=1}^{M} a_k \sum_n s_{n-i}'\, s_{n-k}' \ ,$$

of

$$R_i' = G \sum_n d_n\, s_{n-i}' - \sum_{k=1}^{M} a_k\, R_{i-k}' \ , \tag{3.1.5}$$

waarin $R_i' = \sum_n s_n'\, s_{n-i}'$

de i-de autocorrelatie is van de impulsresponsie van het synthesefilter. Voor i=0 volgt dan uit (3.1.5) voor de totale energie:

$$R_o' = G^2 - \sum_{k=1}^{M} a_k\, R_k' \ , \tag{3.1.6}$$

of:

$$G^2 = \sum_{k=0}^{M} a_k\, R_k' \ , \qquad \text{met } a_o = 1. \tag{3.1.7}$$

En voor $i \neq 0$ volgt uit (3.1.5):

$$R_i' = -\sum_{k=1}^{M} a_k\, R_{i-k}' \ . \tag{3.1.8}$$

Stellen we nu de eis dat de energie R_o' van deze impulsresponsie van het synthesefilter H(z) gelijk moet zijn aan de energie R_o van het oorspronkelijke spraaksignaal in het analysevenster, dan volgt hieruit

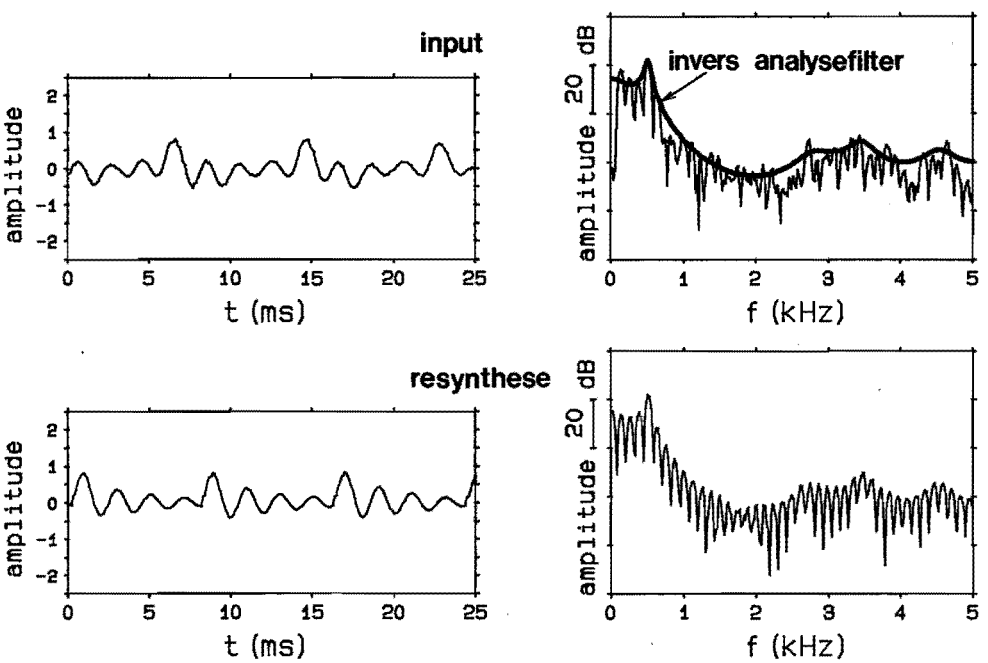

Fig.3.3. Golfvorm en energiespectrum van het ingangssignaal (boven) en de resynthese (onder). De dikke curve rechts boven is het geïnverteerde analysefilter dat bij de gegeven orde M (hier 10) zo goed mogelijk past bij het ingangsspectrum.

---

door vergelijking van (3.1.8) met (2.1.5) dat de M autocorrelaties van de impulsresponsie van H(z) gelijk moeten zijn aan de eerste M autocorrelaties van het oorspronkelijke spraaksignaal (Makhoul, 1975; Markel en Gray, 1976), dus:

$$R_i' = R_i \quad \text{met} \quad i = 0...M \ .$$

Hieruit volgt dan met (3.1.7) en (2.1.6) dat aan deze eis voldaan is als:

(3.1.9) $$G2 = E_m = \sum_{k=0}^{M} a_k R_k \ .$$

Bij de analyse zijn de autocorrelaties /R_k/ en daaruit de filtercoëfficiënten /a_k/ berekend. Door nu tevens het inprodukt van beide uit te rekenen volgens (3.1.9) is daarmee, als 'bijprodukt' van de analyse, de amplitudeversterkingsfaktor G bepaald. Dus als de versterking G, voorafgaand aan het synthesefilter gelijk gemaakt wordt aan de wortel uit de energie van het restsignaal van de analyse, dan is de energie van het geresynthetiseerde signaal gelijk aan die van het oorspronkelijke spraaksignaal in het analysevenster. Dat is hier aange-

toond voor de eenheidsimpuls als excitatie, maar kan precies zo worden bewezen voor ruis als excitatie (Makhoul, 1975).

In het frekwentiedomein betekent dit dat het energiespectrum van de resynthese, dat qua vorm al zo goed mogelijk paste bij het oorspronkelijke spectrum, met de versterkingsfaktor G volgens (3.1.9) nu ook zo goed mogelijk daarmee tot dekking is gebracht, fig. 3.3. Integratie over het gehele frekwentiegebied levert voor de gladde curve hetzelfde op als voor het oorspronkelijke spectrum; beide energieën zijn gelijk.

In de volgende paragraaf zullen we bespreken hoe de resynthese voor complete spraakuitingen in de praktijk wordt uitgevoerd.

3.2 PRAKTISCHE UITVOERING VAN DE RESYNTHESE

In hoofdstuk 2 hebben we gezien hoe bij de analyse de spraaksegmenten op vaste afstanden in de tijd (standaard 10 ms) worden gerepresenteerd door frames, bestaande uit bij elkaar horende parameters voor bronsignaal en analysefilter. Deze opeenvolgende frames zijn opgeslagen in een file in het schijvengeheugen van de computer. Om uit zo'n file met analysedata de spraak weer te synthetiseren worden de opeenvolgende frames teruggelezen van de schijf. Voor ieder frame wordt vervolgens in grote lijnen de bij de analyse gevolgde procedure in omgekeerde richting doorlopen. Dat wil zeggen dat wanneer het analysefilter is beschreven in de vorm van pq-parameters, deze eerst volgens (2.1.20) worden uitvermenigvuldigd tot de a-parameters, de coëfficiënten van het M^e orde analysefilter in direkte vorm. Dan wordt het vaste pre-emfasefilter:

$$(2.3.3) \qquad P(z) = 1 + u\, z^{-1} \quad , \qquad \text{met } u = -.9$$

dat bij de analyse voorafging aan de berekening van het M^e orde analysefilter in rekening gebracht. Daarmee ontstaat een $(M+1)^e$ orde filter:

$$(3.2.2) \qquad B(z) = P(z)\, A(z) = \sum_{k=0}^{M+1} b_k\, z^{-k} \quad ,$$

waarvan de coëfficiënten $/b_k/$ gegeven worden door:

$$(3.2.3) \qquad b_k = a_k + u\, a_{k-1} \quad , \text{ met } a_o=b_o=1 \text{ en } k = 0....M+1.$$

De geïnverteerde van dit filter B(z) is dan het filter waarmee de eigenlijke synthese wordt uitgevoerd. Daartoe wordt de responsie berekend van 1/B(z) op een impuls of witte ruis, conform het produktie-

model van hoofdstuk 1. De uitgangssamples s_n' worden berekend volgens (3.1.4):

$$(3.2.4) \qquad s_n' = G\, u_n - \sum_{k=1}^{M+1} b_k\, s_{n-k}' \quad ,$$

waarin $/b_k/$ de coëfficiënten zijn van het $(M+1)^e$ orde analysefilter inclusief pre-emfase, gegeven door (3.2.3), G de amplitude of versterkingsfaktor is volgens (3.1.9) berekend bij de analyse en u_n het ingangssignaal voor het synthesefilter: een periodieke eenheidsimpuls of witte ruis. De stem/stemloosparameter bepaalt welke van deze twee excitaties wordt toegepast. In stemloze frames is u_n een trekking uit randomgetallen en voor stemhebbende frames is $u_n = 1$ als n overeenkomt met het begin van een grondtoonperiode en anders 0.
Aldus worden voor ieder van de schijf gelezen frame de samples van de resynthese stuk voor stuk berekend. Daarbij zijn twee werkwijzen mogelijk, die we nu kort zullen toelichten.

3.2.1 Asynchrone resynthese

De eerste (en snelste) wijze is die waarbij aan de hand van ieder gelezen frame zoveel samples worden berekend als overeenkomt met de frameperiode, de stapgrootte waarmee de analyse is uitgevoerd. Meestal is die 10 ms, dus bij f_s = 10 kHz worden per frame 100 samples berekend. Alle parameterwaarden worden na afloop van de frameperiode vervangen door de nieuwe waarden die van de schijf zijn gelezen en blijven dan de gehele frameperiode van kracht. Van stemhebbende frames verloopt de grondtoonperiode in het algemeen asynchroon met de frameperiode; er kan overal binnen een frameperiode een excitatie van het filter optreden. Dus kan ook overal binnen een grondtoonperiode overgeschakeld worden op nieuwe filterparameters en zelfs van stemhebbend naar stemloos of omgekeerd, als het nieuwe frame dat oplegt. Overgangen tussen frames zijn dus in het algemeen abrupt.

Daarnaast zijn er, ingeval de frameperiode veel groter is dan de grondtoonperiode, binnen een frame meerdere opeenvolgende grondtoonperiodes waarin de geresynthetiseerde golfvorm exact gelijk is. Zo'n situatie kan zich b.v. voordoen bij vrouwenspraak met hoge toonhoogte, of wanneer bij een zuinige analyse een grote frameperiode is gekozen. In fig. 3.4 is deze situatie weergegeven. We zien in deze asynchrone resynthese abrupte overgangen en bij hoge toonhoogte volstrekt identieke stukken golfvorm binnen een frame. Dit kan onder sommige omstandigheden aanleiding geven tot hoorbare discontinuiteiten in de geresynthetiseerde spraak. Om dat te vermijden en in zo'n situatie de resynthese 'gladder' te laten klinken kunnen we de navolgende, synchrone manier van resynthese toepassen.

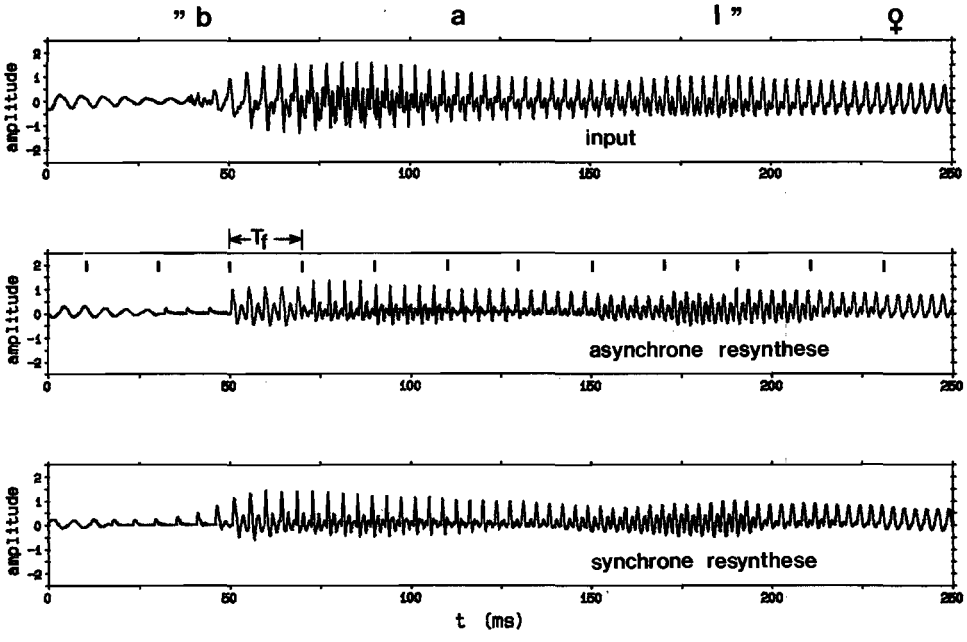

Fig.3.4. Golfvorm van een fragment uit het woordje "bal", uitgesproken door een vrouwenstem, met F_0 van 200 tot 250 Hz. Boven: ingangssignaal. Midden: asynchrone resynthese na analyse met een frameduur T_f van 20 ms (aangegeven door streepjes). Onder: synchrone resynthese met interpolatie tussen opeenvolgende frames. In dit voorbeeld is tevens te zien hoe het korte ruisplofje bij het opheffen van de lipafsluiting bij de b (bovenste curve voor t = 40 ms) in beide gevallen na resynthese niet nauwkeurig bewaard blijft; alle frames zijn als stemhebbend geklassificeerd.

3.2.2 Synchrone resynthese

Bij deze tweede, wat minder simpele methode wordt synchroon met de grondtoonperiode gewerkt. Dat wil zeggen dat aan het begin van iedere grondtoonperiode tegelijk met de excitatie-impuls de parameterwaarden bijgesteld worden. Resultaat van deze werkwijze is weergegeven in fig. 3.4, ook weer voor het geval de grondtoonperiode veel kleiner is dan de frameperiode. Alle parameters, dus ook de waarde van de grondtoonperiode zelf, worden uitsluitend aan het begin van iedere grondtoonperiode vernieuwd, nadat ze berekend zijn als lineaire interpolatie tussen waarden die in de twee dichtstbijzijnde frames zijn opgeslagen. Daardoor verloopt de geresynthetiseerde golfvorm voor de opeenvolgende grondtoonperiodes geleidelijk. Op de grens tussen twee frames wordt de betreffende grondtoonperiode eerst afgemaakt. We zien

in de onderste curve van fig. 3.4 nu geen abrupte overgangen meer zoals die bij synchrone resynthese (middelste curve) wel optreden. Synchrone resynthese zal dus in gevallen waar de frameperiode groot is vergeleken met de periode van de grondtoon minder abrupte overgangen opleveren en 'gladder' klinken dan de asynchrone.

In de praktijk is, zeker bij mannenspraak en een frameperiode van 10 ms, synchrone resynthese niet nodig. De extra rekentijd voor de interpolaties en het feit dat daarvoor ordenen en begrenzen van de pq-parameters volgens par. 2.5, nodig is, wegen dan niet op tegen de niet of nauwelijks hoorbare verschillen met de syntheseresultaten van de asynchrone methode.

Wanneer aldus de samples van een frameperiode (asynchroon) of van een grondtoonperiode (synchroon) zijn berekend worden ze in een nieuwe file op het schijvengeheugen weggeschreven. De laatste M+1 berekende samples, (de inhoud van het filtergeheugen), worden gebruikt bij de berekening van het nieuwe stuk golfvorm van de volgende frame- cq. grondtoonperiode. Dit resultaat wordt op zijn beurt weggeschreven, enz, totdat de gehele file is doorlopen.

Na afloop worden dan alle samples van de geresynthetiseerde golfvorm van de nieuwe file genormeerd op een maximale absolute waarde van $2^{11}-1 = 2047$, zodat het signaal past binnen de 12 bits (11 + tekenbit) van de digitaal-analoogomzetter, via welke de nieuwe file ten gehore kan worden gebracht. Na deze omzetting wordt het analoge signaal vervolgens gefilterd zodat alle komponenten hoger dan de halve samplefrekwentie worden onderdrukt. Dit is dus de omgekeerde weg die bij de spraakinname (par. 2.3.1) is gevolgd.

Daarmee is het gehele analyse-resyntheseproces doorlopen. In de volgende hoofdstukken zullen we de resulaten bespreken die ermee worden verkregen.

4 FYSISCHE EVALUATIE

Alvorens de resultaten van analyse en resynthese te bespreken zullen we in de eerste paragraaf van dit hoofdstuk aangeven door welke beperkende faktoren we verschillen mogen verwachten tussen oorspronkelijke en geresynthetseerde spraaksignalen. Deze beperkingen zijn deels afkomstig van het gehanteerde analysefilter, dat slechts nulpunten heeft en dat de uitgangsenergie minimaliseert over een duur die niet willekeurig klein kan zijn. Deels ook zijn de beperkingen afkomstig van de wijze van resynthese, waarin het filter wordt geëxciteerd met een impulsreeks of met ongecorreleerde ruis, terwijl het restsignaal van de analyse meestal geen echt wit spectrum heeft en bovendien mengvormen van periodieke en ruissignalen kan bevatten.

Tegen de achtergrond van deze beperkingen zal in de tweede paragraaf het geresynthetiseerde signaal worden vergeleken met het origineel. Voor een kunstmatig testsignaal, alsmede voor gewone, natuurlijke spraak, worden golfvorm en energiespectra van het uitgangssignaal vergeleken met die van het ingangssignaal. Uiteengezet wordt hoe de gekonstateerde verschillen in tijd- en frekwentiedomein zijn terug te voeren tot de genoemde beperkingen in het analyse-resynthesesysteem.

Tot besluit wordt dan in de derde paragraaf nagegaan of en in welke mate de vaste parameters in het analyseproces van invloed zijn op de resynthese. Daaruit worden dan voor ieder van deze modelkonstanten de fysisch optimale waarden afgeleid, zoals die bij de standaardanalyse worden toegepast.

4.1. BEPERKENDE FACTOREN

In het vorige hoofdstuk hebben we gezien hoe het spraaksignaal werd geanalyseerd door stap voor stap in de tijd de parameters te berekenen van bronsignaal en analysefilter. Deze parameters hebben we berekend uit een stukje spraaksignaal ter lengte van het analysevenster, bestaande uit N samples. Binnen dit venster wordt het signaal als stationair beschouwd, omdat de filtercoëfficiënten zijn bepaald uit de eerste M autocorrelaties van het signaal, waarbij over alle N samples is gesommeerd. Aangezien autocorrelaties per definitie even funkties zijn van de tijd, levert een in de tijd omgekeerd signaal exact dezelfde filtercoëfficiënten op als het gewone signaal; immers de autocorrelaties zijn in beide gevallen gelijk. Zo kan ook bij de berekening van het analysefilter geen onderscheid worden gemaakt tussen toe- en afnemende amplitudes. Het energiespectrum en daarmee het analysefilter zijn dus ahw gemiddelden voor het gehele signaal binnen het analysevenster. In theorie is de analyse dus alleen geldig voor (lokaal) stationaire signalen. Dit houdt in dat de lengte van het analysevenster in principe af zou moeten hangen van het ingangssignaal. Voor een klinker met een laagste grondtoon van bv 50 Hz zou het venster minstens 40 ms moeten zijn om aan de eis van stationair zijn enigszins te voldoen. Maar zo'n vensterlengte van 40 ms is voor snelle overgangen, zoals die onder meer bij plofklanken optreden, in principe weer te groot. Deze overgangen spelen zich soms binnen 10 ms of minder af en worden dan in de analyse uitgesmeerd over de duur van het venster. In theorie zou dus een variabele lengte van het analysevenster nodig zijn, voor ieder frame aangepast aan de lokale situatie. Tijdens de analyse zou dan vastgesteld moeten worden of het spraaksignaal zich in een (kwasi-) stationaire dan wel in een overgangsfase bevindt. Helaas is het in de praktijk zeer moeilijk om zoiets uit te voeren; goede criteria op grond waarvan zo'n onderscheid gemaakt kan worden ontbreken. In onze standaardanalyse beperken we ons dan ook tot een vaste lengte van het analysevenster met een compromiswaarde van 25 ms. In tegenstelling tot kwasi-stationaire stukken (klinkers) zullen snelle overgangen in het spraaksignaal (o.m. plofklanken) in principe dan ook minder goed worden weergegeven. Van deze eerste beperking zullen we in de volgende paragraaf de gevolgen zien.

Een tweede beperking die we ons bij de analyse hebben opgelegd ligt in het feit dat het analysefilter uit louter nulpunten bestaat. Spraaksegmenten waarvan het spectrum al nulpunten bevat (nasalen) worden daarom door zo'n analysefilter met een beperkt aantal nulpunten niet optimaal geneutraliseerd. Om die nulpunten in het ingangsspectrum effektief te kunnen 'uitvlakken' zou het analysefilter ook polen moeten bevatten. Een analysefilter met polen levert echter niet-

lineaire vergelijkingen op die iteratief opgelost moeten worden (Makhoul, 1975). Dat zou dan tot gevolg hebben dat de coëfficiënten van zo'n (met polen uitgebreid) filter niet meer snel en eenvoudig uit het spraaksignaal zelf berekend kunnen worden, een van de eisen die we aan het systeem hebben gesteld. In onze analysemethode zien we daar dan ook van af.

Niet alleen aan de analysekant, ook aan de synthesekant zijn een tweetal beperkingen van belang.

Een daarvan, de derde beperking, houdt in dat we het synthesefilter exciteren met impulsen of met ongecorreleerde (witte) ruis. Hoewel het restsignaal van de analyse doorgaans noch deze ideale impulsvorm heeft, noch ongecorreleerde ruis is, wordt het synthesefilter toch met zo'n geidealiseerd signaal aangeslagen. Daardoor brengen we al een belangrijke reduktie van de datastroom aan. Immers het restsignaal bevat naast alle faseinformatie van het oorspronkelijke spraaksignaal ook dat deel van het energiespectrum dat door het filter niet geneutraliseerd kan worden. Door in plaats van dit restsignaal (dat niet berekend wordt in de gewone analyse) een impulsreeks of witte ruis als bronsignaal te nemen leggen we ons een beperking op die twee belangrijke gevolgen heeft.

Ten eerste zal in het frekwentiedomein, althans voor gewone spraaksignalen, het energiespectrum na resynthese afwijken van het ingangsspectrum. Die afwijkingen zullen kleiner zijn naarmate er meer coëfficiënten worden gebruikt voor het analysefilter. Het spectrum van het restsignaal zal dan meer gaan lijken op het witte spectrum van impulsen of witte ruis. Maar we willen met ons systeem juist spraaksignalen met zo weinig mogelijk parameters beschrijven en we zullen in de komende paragraaf en hoofdstukken laten zien dat een tiental coëfficiënten voldoende is. Op grond van dit beperkte aantal kunnen we dan wel verschillen verwachten tussen de omhullende van de energiespectra van input en output van het systeem.

Het tweede gevolg van het 'wit' exciteren van het synthesefilter is dat daarmee ook de faseinformatie van het ingangssignaal verloren gaat. Zelfs al zou met een groot aantal coëfficiënten worden geanalyseerd zodat de omhullende van het energiespectrum exact zou worden beschreven door het synthesefilter, dan nog zou in het tijddomein de golfvorm er na resynthese anders uitzien dan die van het oorspronkelijke spraaksignaal. Analyse- en synthesefilter zelf zijn minimumfase, dwz dat ze alleen de fasedraaiingen bevatten die minimaal nodig zijn om een bepaald energiespectrum te realiseren. Dat betekent dat de golfvorm na resynthese alleen exact overeenkomt met de oorspronkelijke wanneer het ingangssignaal zelf al een minimumfase signaal is en tevens de spectrale omhullende exact door het filter wordt beschreven. We zullen daarvan in de volgende paragraaf een voorbeeld zien

voor kunstmatig gegenereerde testsignalen. Bij gewone spraak moeten we echter op grote verschillen rekenen in het tijddomein, tengevolge van deze derde beperking.

De vierde beperking is de uitsluiting van een kombinatie van stemhebbende en stemloze klanken bij de resynthese. Excitatie van het synthesefilter gebeurt òf met impulsen, òf met witte ruis. Klanken waarin zowel periodieke als (kort durende) ruiscomponenten aanwezig zijn, zullen daardoor fysisch worden aangetast. Dat is niet alleen te verwachten bij stemhebbende wrijfklanken maar ook bij bv. stemhebbende plofklanken. We zullen hierop in hoofdstuk 5 nog uitvoerig terugkomen.

Tegen de achtergrond van deze beperkende faktoren zullen we in de volgende paragraaf input en resynthese met elkaar vergelijken.

4.2. VERGELIJKING TUSSEN INPUT EN OUTPUT VAN HET SYSTEEM

In deze paragraaf zullen we de resultaten van analyse en resynthese bespreken, in hoofdzaak door het ingangssignaal te vergelijken met het uitgangssignaal, zowel in tijd- als in frekwentiedomein. Deze fysische evaluatie is sterk visueel gericht, door vergelijking van plaatjes van de beide golfvormen, energiespectra en resogrammen. Uiteindelijk gaat het er natuurlijk om hoe de spraak na resynthese klinkt en daarvoor is een perceptieve evaluatie nodig, die in hoofdstuk 5 aan de orde komt. Met de plaatjes die hier worden gegeven kunnen we evenwel de belangrijkste eigenschappen van het systeem illustreren.

Analyse en resynthese zijn uitgevoerd op standaard wijze, zoals in hoofdstuk 2 en 3 is beschreven. Het ingangssignaal is bemonsterd op 10 kHz en is na toepassing van een 25 ms Hammingwindow en vaste preemfase van u = -0.9 geanalyseerd en vervolgens synchroon geresynthetiseerd. De grondtoonfrekwentie van de stemhebbende stukken van de kunstmatig gegenereerde testsignalen heeft een vaste waarde van 100 Hz. Bij de natuurlijke spraak zijn na afloop van de toonhoogtemeting enkele oktaaf- en stemloosfouten met de hand gecorrigeerd.

Eerst zullen we de resultaten bespreken van het kunstmatige testsignaal, daarna voor spraak als ingangssignaal.

4.2.1. Kunstmatig ingangssignaal

Als input voor het analyse-resyntheseproces is een kunstmatig gegenereerd spraakachtig signaal gebruikt. Dit bestaat uit een stukje klinker van 50 ms, dan 70 ms ruis en weer de klinker, gevolgd door 25 ms stilte en tot besluit een ruisstootje van 10 ms (fig. 4.1, bovenste curve). Dit signaal is tot stand gekomen door filterparameters te specificeren overeenkomstig een zgn neutrale klinker, (resonanties

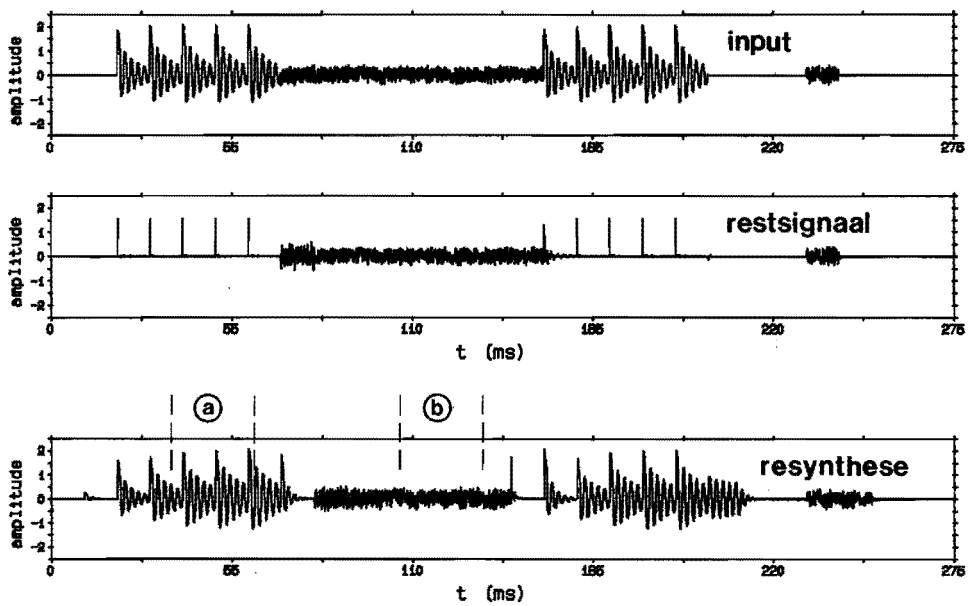

Fig.4.1. Boven: golfvorm van het kunstmatig opgewekte testsignaal, gebruikt als input voor het systeem. Onder: golfvorm van de resynthese. Midden: restsignaal van de analyse. Van de segmenten (a) en (b) zijn de spectra van input en resynthese weergegeven in fig. 4.2 .

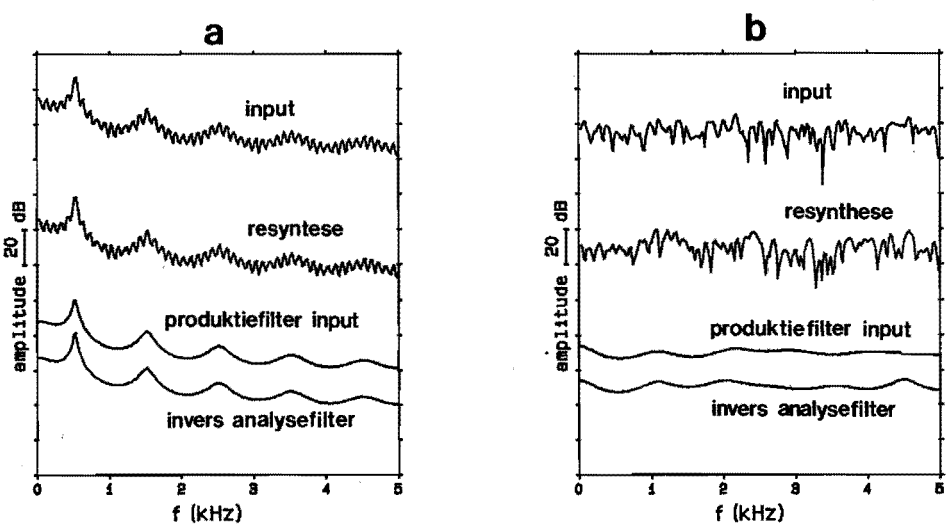

Fig.4.2. Spectra van de stationaire segmenten (a) en (b) uit fig. 4.1 voor stukjes input (bovenste curves) en resynthese (70 db daaronder), links voor de klinker, rechts voor ruis. De gladde curves zijn de spectra van het filter waarmee de input is gegenereerd en (20 dB verschoven) van het berekende geïnverteerde analysefilter.

op (2n-1)*500 Hz, n=1 t/m 5, en een konstante kwaliteitsfaktor van 10), ze op te slaan in een file en deze dan op de gewone wijze te synthetiseren. Dit kunstmatige spraaksignaal (de input) is vervolgens geanalyseerd en geresynthetiseerd, beide volgens de standaardmethode. Het resultaat daarvan (de output) is in fig. 4.1, onderste curve weergegeven. Fig. 4.2 toont de spectra van twee stationaire fragmenten van 25 ms uit zowel klinker als ruis. We konstateren het volgende:

- In de stationaire periodieke stukken zijn zowel de golfvorm als het energiespectrum van input en resynthese identiek. Voor dit synthetische spraaksegment is dat in overeenstemming met wat in par. 4.1 is opgemerkt: het ingangssignaal voor de analyse is een minimumfasesignaal, waarvoor bovendien de omhullende door het analysefilter perfekt wordt vlakgestreken. Het is immers gegenereerd door de inverse van het analysefilter. Geheel volgens verwachting is het restsignaal van de analyse blijkens fig. 4.1 (middelste curve), een perfekte reeks impulsen.
- Ook in de stationaire ruisstukken vertonen spectrale omhullende van input en output goede gelijkenis, zoals uit fig. 4.2 blijkt door vergelijking tussen de energiespectra van het synthesefilter dat bij het genereren van het betreffende stukje input is gebruikt en het spectrum van het geinverteerde analysefilter. We zien alleen boven 2 kHz wat kleine verschillen die zijn toe te schrijven aan het feit dat (voor dit stukje van het testsignaal) het inputspectrum geproduceerd is met een stochastisch bronsignaal.
- Met betrekking tot de overgangen tussen stilte en signaal zien we in fig. 4.1 (onderste curve) duidelijk hoe deze worden uitgesmeerd en verlengd over de duur van het analysevenster, in dit geval dus ongeveer 2 frames. Bij het begin van het signaal start het signaal een frame te vroeg vergeleken met het ingangssignaal en aan het einde gaat het te lang door. Hetzelfde beeld zien we voor het ruisstootje aan het einde; ook dat wordt velengd tot ongeveer 25 ms.
- Bij de overgangen tussen klinker en ruissignaal blijkt iets dergelijks met betrekking tot de stem/stemloosdetector. Deze klassificeert frames ten onrechte als stemhebbend in die gevallen waarin de ene helft van het signaal binnen het analysevenster nog (alweer) een stuk klinker bevat en de andere helft al (nog) ruis.

Concluderend zien we voor dit kunstmatige testsignaal dat de stationaire stukken perfekt door de analyse en de resynthese heen komen. Niet-stationaire stukken komen niet ongeschonden door de analyse, ze worden in het tijddomein aangetast. Deze fout hangt samen met het voor deze overgangen te lang gekozen analysevenster en ze zou verkleind kunnen worden door dat venster korter te kiezen. Dat heeft echter, zoals we in par. 4.3 zullen zien, tot gevolg dat daarmee dan juist de

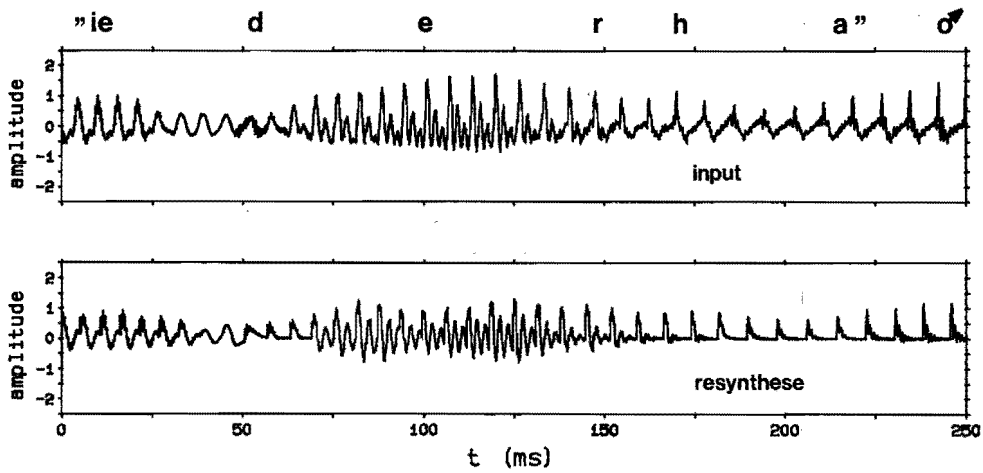

Fig.4.3. Voorbeeld van de golfvorm van input (boven) en resynthese (onder) van een fragment mannenspraak uit "ieder half..".

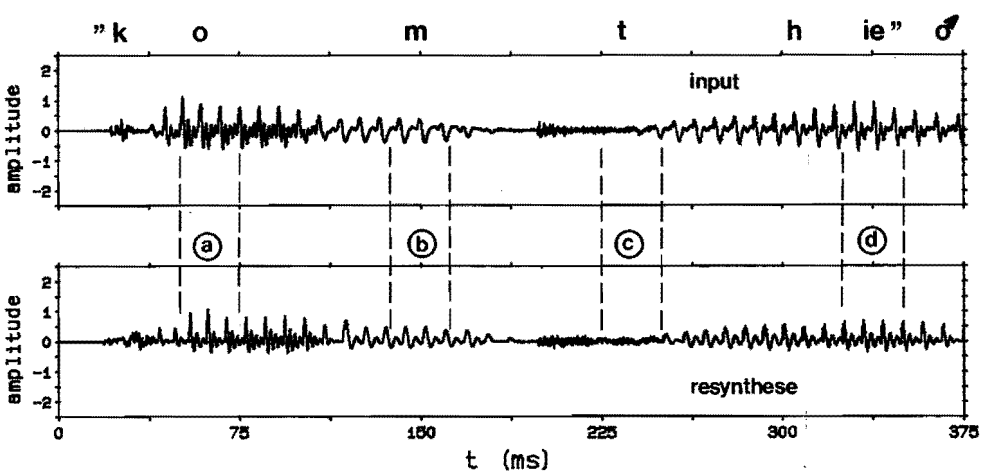

Fig.4.4. Als fig. 4.3 maar nu voor het fragment "..komt hier..". Van de segmenten (a) t/m (d) zijn in fig. 4.5 de energiespectra van input en resynthese weergegeven.

stationaire stukken weer slechter worden weergegeven.

Gelukkig betekent deze fout nog niet dat daarmee ook bij natuurlijke spraak de overgangen zo slecht door het analyseproces worden beschreven. Daarin komen de zeer scherpe discontinuïteiten, die hier kunstmatig in het ingangssignaal zijn aangebracht, niet in die mate voor. In de volgende paragraaf zullen we zien hoe deze spraak door het systeem heen komt.

4.2.2. Natuurlijke spraak als ingangssignaal

Resultaten van de standaard-analyse en resynthese zijn weergegeven in fig. 4.3 en 4.4 voor delen van de zin "ieder half uur komt hier een bus langs", uitgesproken door een mannnenstem. Kwasi-stationaire fragmenten zijn te zien in fig. 4.3 en een stuk met relatief snelle overgangen in fig. 4.4. De bovenste plaatjes zijn steeds de golfvormen van de inputspraak, de onderste die van de resynthese.

We konstateren in het tijddomein overeenkomsten en verschillen. Overeenkomsten als we kijken naar het globale verloop van amplitude en grondtoonperiode. Verschillen wanneer we details in de golfvorm bezien.

- Als eerste verschil zien we hoe in fig. 4.3 na de resynthese de gemiddelde waarde van het ingangssignaal positief is, in tegenstelling tot de input, waarvan het gemiddelde (de DC component) nul is gemaakt. Dit komt omdat de gemiddelde waarde van het ingangssignaal voor het synthesefilter eveneens positief is; de excitatieimpuls voor het filter heeft de amplitude G of nul. Dit verschil heeft verder voor ons geen consequenties.
- Als tweede zien we vooral in fig. 4.3 tussen t = 150 en 250 ms dat de golfvorm na resynthese veel pulsvormiger is geworden. De energie zit na resynthese meer geconcentreerd aan het begin van iedere periode. Dit verschil hangt samen met het in par. 4.1. besproken feit dat bij resynthese een impuls als excitatie is gebruikt, waarmee dus de faseinformatie van het ingangssignaal verloren is gegaan. De resynthese is een minimumfase-signaal geworden.
- Als derde verschil tussen input en output zien we hoe bij snelle overgangen (de ruisstukjes bij de k en de t van "komt") het signaal iets wordt uitgesmeerd in de tijd doordat het analysevenster 25 ms lang is. Zoals we in par. 4.1 hebben gezien is dit eigenlijk te lang voor snelle overgangsverschijnselen. Verder heeft dit tot gevolg dat het korte ruisplofje bij de d van "ieder" (bovenste curve in fig. 4.2 bij t = 50 ms) niet gedetecteerd wordt door de stem/stemloosdetector en na resynthese dan ook verdwenen is. Voor de plofklank b hebben we al eerder in fig. 3.4 een soortgelijke aantasting gezien.

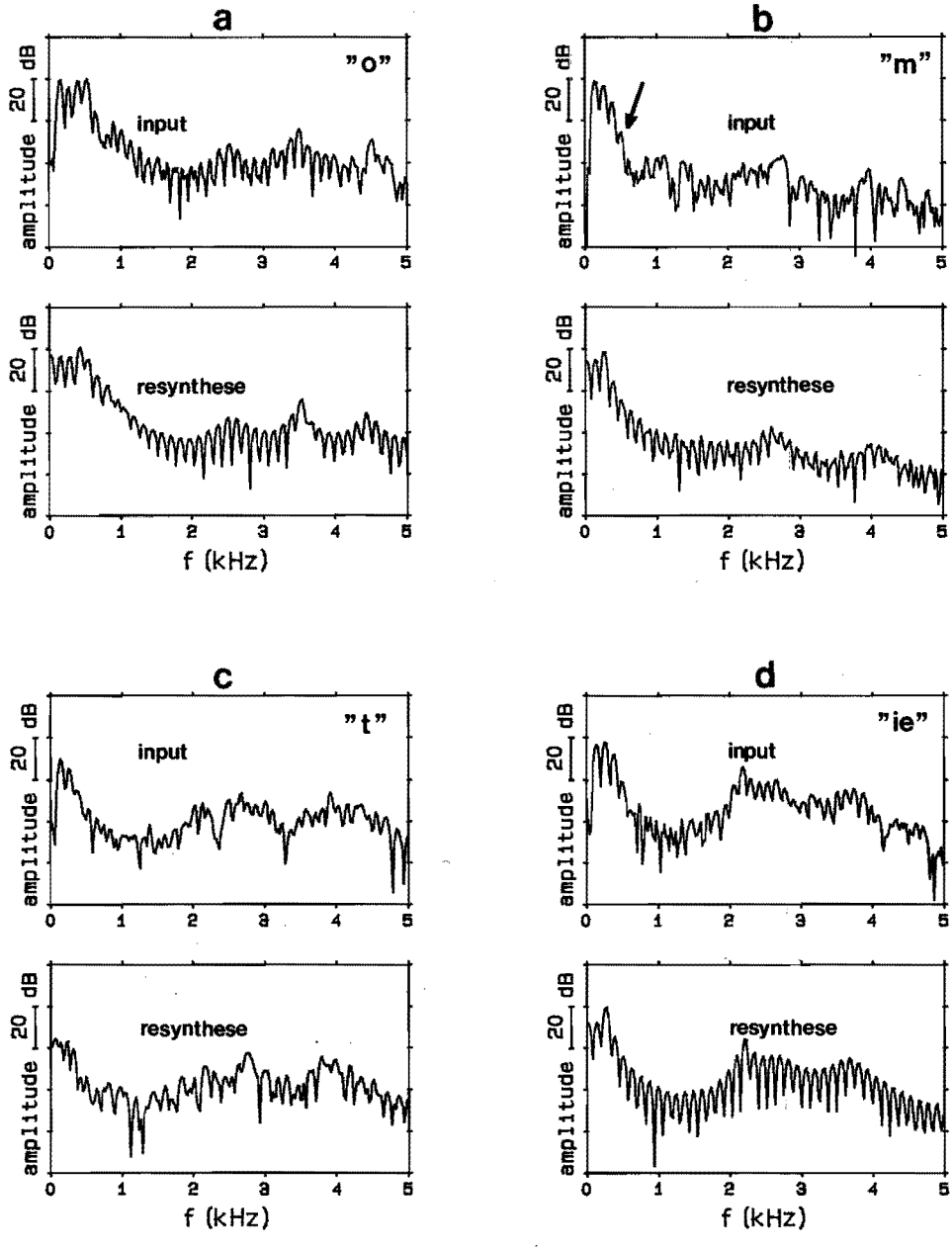

Fig.4.5. Energiespectra van input (bovenste plaatjes) en resynthese (onderste plaatjes) voor de segmenten (a) t/m (d) uit fig. 4.4 van resp. de o, m en t uit "komt" en de ie uit "hier".

Ook in het frekwentiedomein konstateren we zowel verschillen als overeenkomsten tussen de energiespectra van output en input. Van de segmenten a t/m d uit fig. 4.4 zijn de bijbehorende spectra weergegeven in fig. 4.5 a t/m d. In ieder plaatje staat boven steeds de input, onder de resynthese. Zoals te verwachten is vertoont het globale verloop van beide spectra overeenkomsten, maar laat een gedetailleerde beschouwing verschillen zien.

- Als eerste noemen we het verschil in de energiespectra voor frekwentie f = 0, bij de stemhebbende fragmenten a, b en d. Deze DC-komponent is na resynthese niet meer 0, zoals reeds besproken.
- Verder zien we verschillen tussen de fijnstruktuur van input en resynthese. In het algemeen vertoont de fijnstruktuur van de resynthese bij stemhebbende klanken (fig. 4.5a, b en d, onderste curves) een veel regelmatiger beeld dan het originele spectrum. Dat is ook te verwachten, immers voor het resynthesefilter wordt een reeks impulsen als excitatie gebruikt en dat is een regelmatiger signaal dan het restsignaal van de analyse van natuurlijke spraak. In de fijnstruktuur van het outputspectrum komt de herhalingsfrekwentie van de impulsen overeen met de afstand tussen twee opeenvolgende pieken, harmonischen van de grondtoon F_o. Bij stemloze stukken, zoals in fig. 4.5c, zijn de fijnstrukturen van input en outputspectrum natuurlijk niet gelijk vanwege de stochastische excitatie van het resynthesefilter.
- Als derde verschilpunt konstateren we hoe soms ook de omhullende van het outputspectrum zichtbaar afwijkt van die van de input. Een voorbeeld daarvan is de m van "komt", het segment van fig. 4.5b, waar in het ingangsspectrum een steile flank ligt tussen 400 en 700 Hz (aangegeven door de pijl) en die na resynthese veel minder steil is geworden. We zien hier een voorbeeld van een nasale klank waarbij met het beperkte aantal van 10 filtercoëfficiënten van het analysefilter zo'n nulpunt niet goed geneutraliseerd kan worden. Daarvoor zou het analysefilter ofwel polen moeten bevatten ofwel veel meer nulpunten.

Concluderend zien we dat voor natuurlijke spraak er nog verschillen kunnen optreden tussen input en resynthese. Deze worden veroorzaakt doordat:

- De analyse wordt uitgevoerd met een vrij lang tijdvenster, waardoor snelle overgangen in het spraaksignaal worden aangetast.
- Het analysefilter uitsluitend nulpunten bevat en dus in principe niet volledig is uitgerust om spectra te beschrijven die zelf al nulpunten bevatten zoals bij nasalen.
- De fase-informatie van het ingangssignaal verloren gaat. Daarmee wordt de fijnstruktuur van het geresynthetiseerde energiespectrum

van stemhebbende signalen veel regelmatiger van struktuur dan het oorspronkelijke spectrum.

Na het vaststellen van deze fysische verschillen tengevolge van beperkingen in het analyse-resynthese-systeem is de vraag in hoeverre deze beperkingen ook leiden tot verschillen in perceptie van de oorspronkelijke en de geresynthetiseerde spraak. Alleen uit de plaatjes is dat zeker niet af te leiden. Die kunnen hoogstens een richting aangeven waarin mogelijke perceptieve verschillen te verwachten zijn. Verschillen die in plaatjes groot lijken kunnen in luisterproeven soms niet of nauwelijks hoorbaar blijken. Omgekeerd is een goede visuele overeenkomst nog geen garantie voor een grote perceptieve overeenkomst. In hoofdstuk 5 zullen dan ook de perceptieve gevolgen van deze beperkingen aan de orde komen, maar eerst gaan we in de volgende paragraaf na welke fysische invloed de gekozen waarden van de vaste modelparameters hebben op de resultaten van analyse en resynthese.

4.3 INVLOED VAN DE VASTE MODELPARAMETERS

De voorbewerkingen die in par. 2.3 zijn genoemd bij de spraakinname in de computer worden standaard toegepast bij digitale spraakbewerking en behoeven hier geen nadere diskussie. Dat ligt anders met de keuze van de modelkonstanten: de lengte L_w (in ms) van het analysevenster (bestaande uit N samples), de pre-emfasekonstante u, het aantal filtercoëfficiënten M en de frameperiode T_f (in ms). Hiervoor hebben we in par. 2.3 standaard-waarden genoemd, zonder daarvan een nadere motivering te geven. Daarom zal in deze paragraaf voor ieder van deze konstanten worden nagegaan wat hun fysische invloed is op de analyseresultaten. Op grond daarvan wordt dan een nadere motivering gegeven voor de waarden die we in ons systeem standaard toepassen.

4.3.1. Invloed van de vensterlengte L_w

In par. 2.3 en 4.2 hebben we als standaardlengte voor het analysevenster 25 ms genomen. Daarbij is in par. 4.2 betoogd dat het venster enerzijds minstens een grondtoonperiode van het spraaksignaal moet bevatten om aan de voorwaarde van lokaal stationair zijn enigszins te voldoen. Anderzijds zal vanwege het uitsmeereffekt de vensterlengte niet te groot mogen zijn omdat dan signalen met snelle overgangen (zoals plofklanken) te zeer aangetast zouden worden. Hier zullen we voor beide reeds eerder toegepaste ingangssignalen nagaan wat het analyseresultaat is voor een zeer kort venster van 10 ms en een zeer lang van 60 ms.

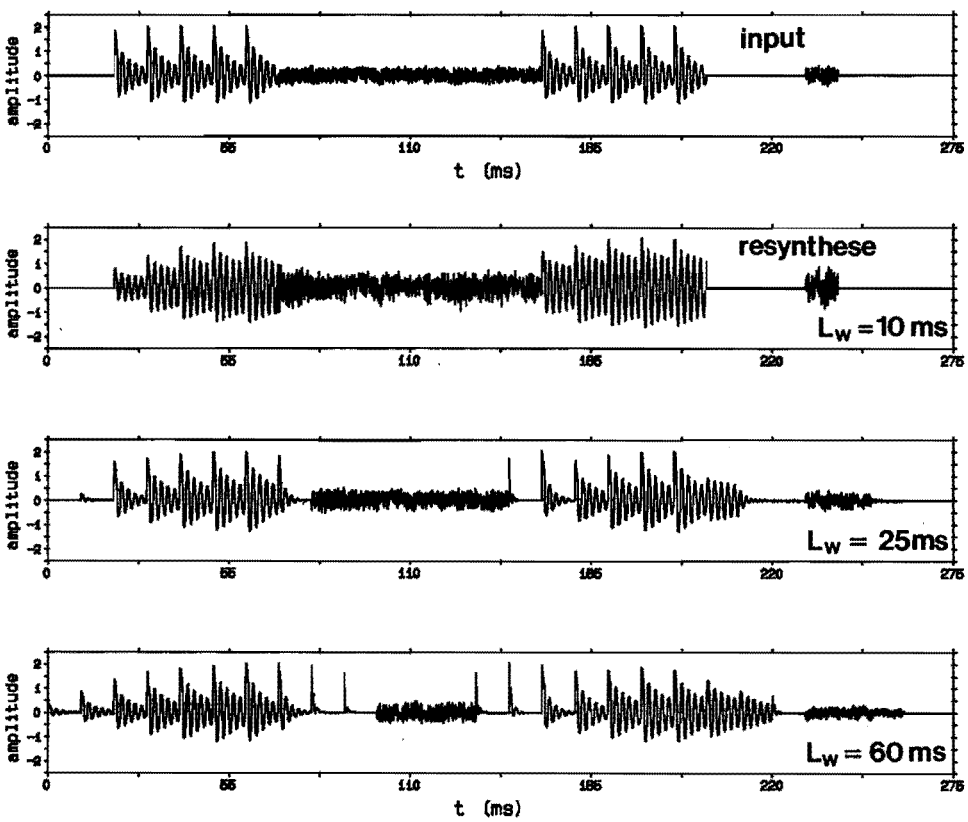

Fig.4.6. Golfvormen van het kunstmatige tessignaal (boven) met daaronder die van de resynthese na analyse met resp. een kort venster van 10 ms, een standaardvenster van 25 ms en een lang venster van 60 ms.

Voor het kunstmatige testsignaal zijn in fig. 4.6 van boven naar beneden de golfvorm van input en resynthese met L_w = 10, 25 en 60 ms onderling te vergelijken. Bij het venster van 10 ms blijft de temporele struktuur perfekt bewaard, zoals we zien aan de correcte postitie van de overgangen tussen signaal en stilte en tussen periodiek en ruissignaal en de stem/stemloosdetektor vertoont nu geen fouten meer bij deze overgangen. Maar in de stationaire stukken van de klinker ontstaan na resynthese grote verschillen met de input. De bandbreedte van de laagste resonantie is veel te smal geworden waardoor binnen een grondtoonperiode het signaal na resynthese te weinig gedempt is, zoals vergelijking tussen de twee bovenste curves nabij t = 55 ms illustreert. Voor dit korte analysevenster voldoet het signaal niet aan de voorwaarde van stationair zijn en is het analysefilter dan ook niet

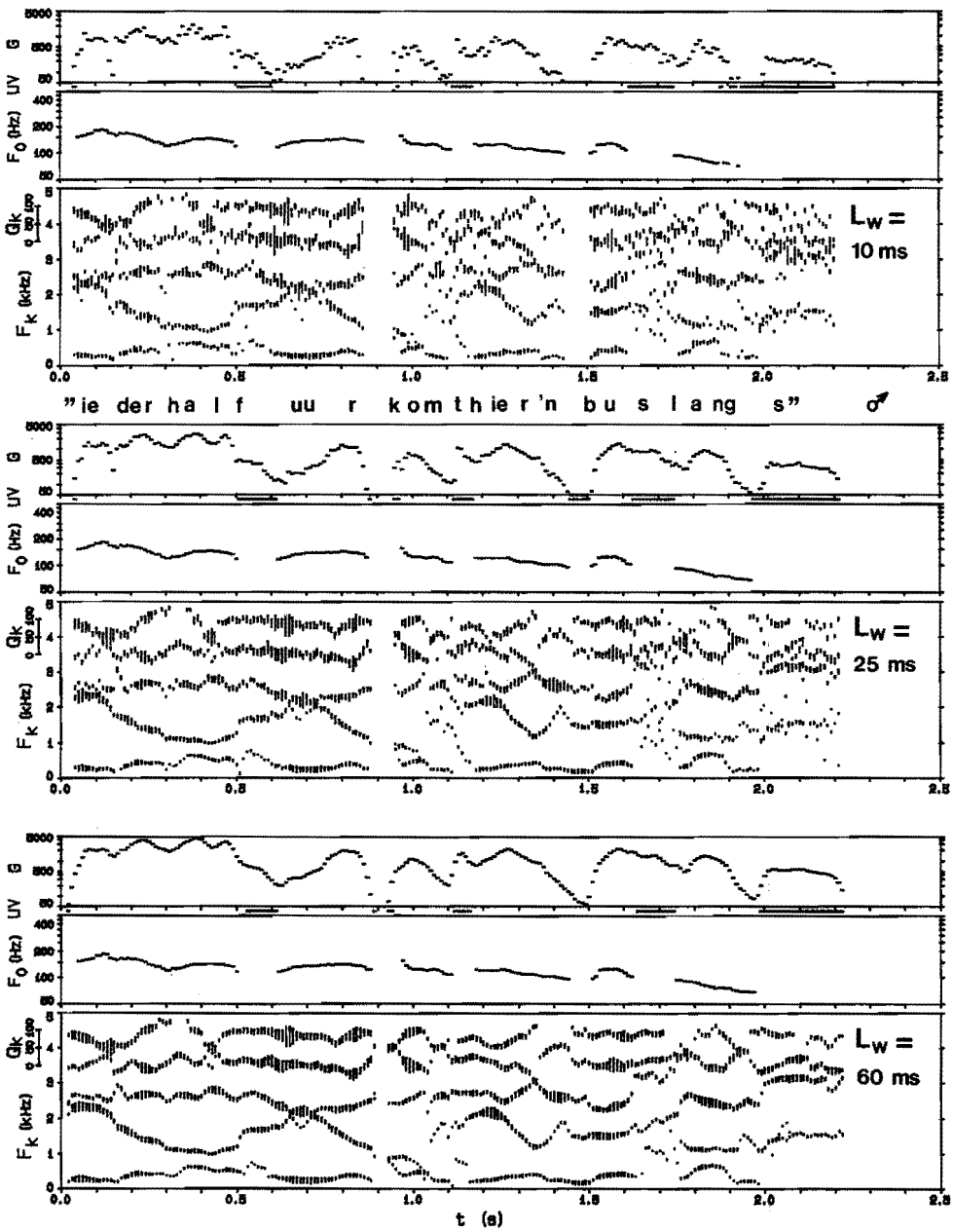

Fig.4.7. Analyseresultaten voor resp. een kort venster van 10 ms (boven), het standaardvenster van 25 ms (midden) en een lang venster van 60 ms (onder).

goed in staat om het spectrum enigszins vlak te strijken.

Verder zal uit fig. 4.6 duidelijk zijn dat een lang venster van 60 ms de tijdstruktuur van het signaal te zeer aantast. De overgangen worden nu veel te sterk uitgesmeerd en de ruis tussen beide klinkers wordt te kort als stemloos geklassificeerd. Tussen t = 70 en 100 ms en tussen t = 130 en 150 ms is de resynthese ten onrechte periodiek geworden.

Voor gewone natuurlijke spraak zijn in fig. 4.7 voor vensters van 10, 25 en 60 ms de analyseresultaten weergegeven voor de zin "ieder half uur komt hier 'n bus langs". Voor het 10 ms venster vertonen zowel amplitude als filterparameters een grillig verloop in de tijd. De analyseresultaten voor de opeenvolgende frames hangen sterk af van toevallige posities van de grondtoonperiode binnen het analysevenster. Voor het 60 ms venster daarentegen verandert met name de amplitude van de opeenvolgende frames te langzaam.

In fig. 4.8 is voor het fragment "komt hier 'n bus" de golfvorm uitgezet. Van boven naar beneden: de input en de resynthese na analyse met een kort, standaard en lang venster van resp. 10, 25 en 60 ms. Bij het korte venster van 10 ms is een aantal frames bij de b van "bus" (aangegeven met de pijl) ten onrechte nul worden. In het algemeen verloopt de resynthese veel 'rafeliger' dan bij 25 ms en klinkt ook ruwer. Verder zien we hoe bij het lange venster de overgangen tussen stemhebbend en stemloos bij de k, de t en de s duidelijk zijn aangetast.

Hoewel het, zoals eerder gezegd, uiteindelijk niet alleen om de visuele effekten gaat, geven de hier gepresenteerde plaatjes wel een indicatie van wat we auditief mogen verwachten. Deze verwachtingen worden ook door luisterproeven met gewone spraak bevestigd. Een kort venster van 10 ms blijkt vooral in de klinkergedeeltes van de geresynthetiseerde spraak aanleiding te zijn tot duidelijk hoorbare 'riedels'; de spraak klinkt ruwer dan na analyse met een lang venster. Omgekeerd levert een lang venster bij de plofklanken een duidelijk slechtere weergave. We behandelen dit verder in hoofdstuk 5.

De conclusie die we uit deze resultaten trekken is dat een variabele vensterlengte, aangepast aan de lokale situatie, theoretisch de voorkeur heeft. In de praktijk is het echter moeilijk om goede kriteria te vinden waarop de vensterlengte gebaseerd kan worden. Daarom passen we in onze analyse als compromis een vensterlengte toe van 25 ms, dus N = 250 bij 10 kHz samplefrekwentie. Voor de langste grondtoonperiode van 20 ms (F_o = 50 Hz) is dan nog een hele periode in het analysevenster aanwezig, terwijl de plofklanken nog voldoende door het systeem worden

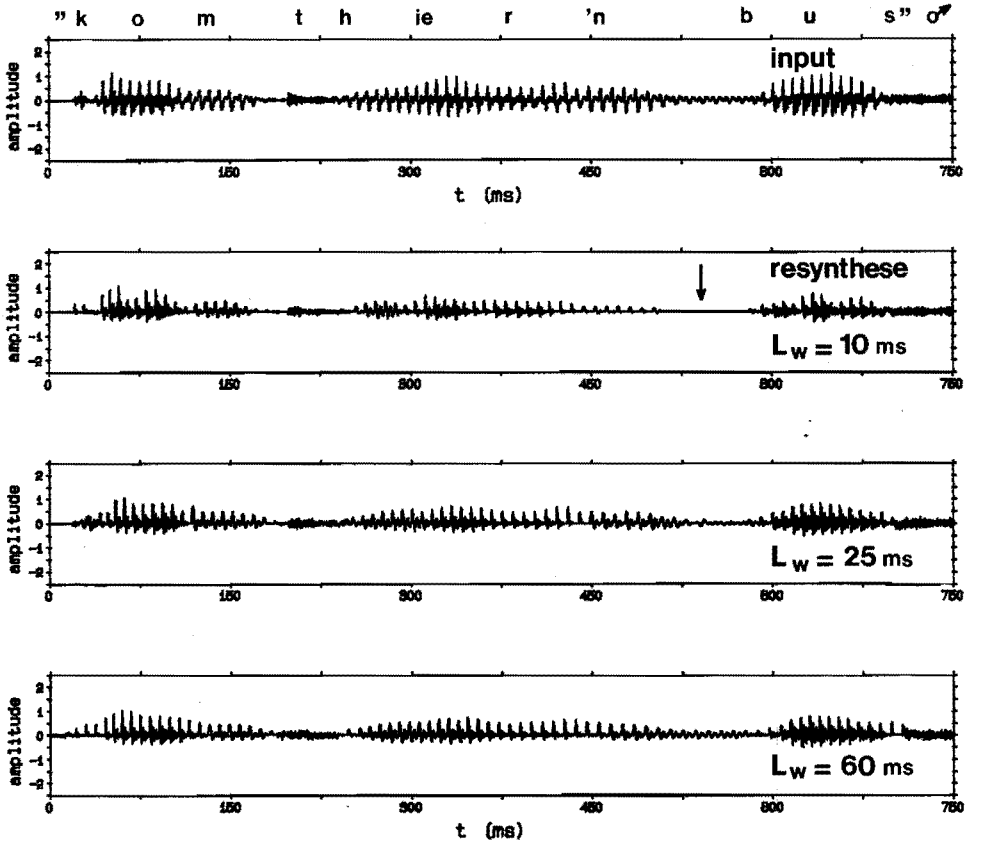

Fig.4.8. Golfvorm van inputspraak (boven) met daaronder die van de resynthese na analyse met resp. een kort venster van 10 ms, het standaardvenster van 25 ms en een lang venster van 60 ms.

---

weergegeven, zoals zal blijken uit de perceptieve evaluatie in hoofdstuk 5.

4.3.2. Invloed van de pre-emfasekonstante u

Bij de standaardanalyse wordt een vaste pre-emfase toegepast van 0.9 waardoor in het spectrum van het ingangssignaal de hoge frekwenties relatief versterkt worden t.o.v. de lage. Deze vaste pre-emfase is in par. 2.3 gemotiveerd met de constatering dat in het algemeen de globale helling van het energiespectrum, gemiddeld over langere tijd, ongeveer -6 dB/oct bedraagt. Voor stemhebbende stukken is dat het gezamenlijke effekt van het stembandspectrum (-12 dB/oct) en het stra-

lingseffekt (+6 dB/oct) aan de mondopening. Voor stemloze stukken vervalt de eerste, zodat daar meestal een globale helling resulteert van +6 dB/oct. Echter over langere spraakuitingen als geheel genomen zijn stemloze stukken doorgaans ver in de minderheid, zowel wat betreft hun totale duur als hun amplitude.

Om deze reden is het voor de handliggend deze vaste helling vóór de analyse begint alvast zoveel mogelijk vlak te maken. Daarmee komt a.h.w. een filtercoëfficiënt vrij voor de eigenlijke analyse. Welk effekt zo'n pre-emfase heeft op de analyseresultaten is nagegaan voor het al eerder gebruikte kunstmatige testsignaal en voor natuurlijke spraak als ingangssignaal.

Voor het kunstmatige testsignaal zien we dit geïllustreerd in fig. 4.9, links voor een stemhebbend frame uit de klinker en rechts voor een frame uit de ruis. Van boven naar beneden zijn achtereenvolgens uitgezet: de spectrale omhullende waarmee het ingangssignaal is gegenereerd, het geinverteerde analysefilter met pre-emfase van resp. -0.9, 0 en +0.9 en tenslotte het ingangsspectrum zelf.

We zien hoe voor de klinker, geanalyseerd met preemfase -0.9, over

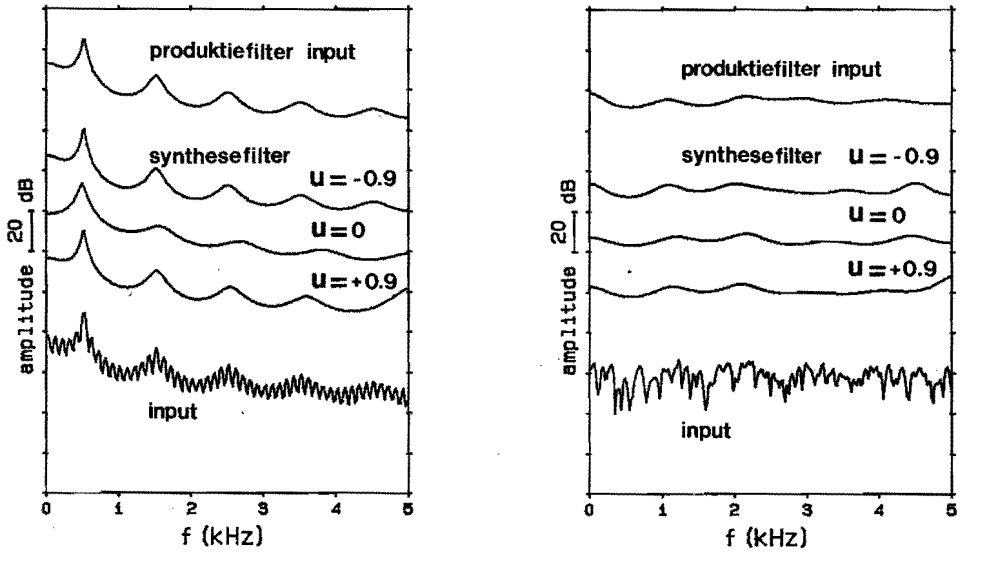

Fig.4.9. Energiespectra, links voor de klinker als testsignaal, rechts voor ruis. Bovenste curves: het filter waarmee de input is gegenereerd. Daaronder (steeds 20 dB verschoven) de geinverteerde analysefilters berekend na pre-emfase van resp. -0.9, 0 en +0.9. De onderste curves zijn (40 dB verschoven) de inputspectra zelf.

eenkomend met een helling van ongeveer +6 dB/oct., het analysefilter goed overeenstemt met het ingangsspectrum. Voor u = 0 ontstaan afwijkingen in de hoger gelegen resonanties. Een preemfase van +0.9 leidt tot een te kleine bandbreedte bij de 500 Hz resonantie en ook tot afwijkingen bij de hogere frekwenties, waarbij de 4.5 kHz resonantie geheel is verdwenen.

Voor het stemloze frame zien we dat bij pre-emfase 0 de omhullende van het ingangsspectrum het best wordt benaderd.

Voor natuurlijke spraak is een tweetal frames weergegeven in fig. 4.10, met van boven naar beneden: preemfase u = -0.9, 0, +0.9 en het ingangsspectrum zelf. Bij het stemhebbende frame (links) levert u = -0.9 de beste passing, maar bij het stemloze frame is een pre-emfase van 0 beter.

Uit beide resultaten valt af te leiden dat aanpassing van de preemfase aan de globale helling van het ingangsspectrum in principe tot de beste analyseresultaten zal leiden. Dus u zou in theorie variabel moeten zijn: voor ieder frame afzonderlijk een waarde tussen -1 en +1, gegeven door de verhouding R_1/R_0 van het signaal binnen het analysevenster. Dat komt dan echter in feite neer op een uitbreiding van het analysefilter met een extra coëfficiënt en we willen met dat aantal juist zo zuinig mogelijk zijn. Een tweede mogelijkheid is dan nog om twee vaste waarden toe te passen, bv. -0.9 voor de stemhebbende en 0 voor de stemloze frames. Het bezwaar daarvan is echter dat in de praktijk eventuele fouten van de stem/stemloosdetector niet meer achteraf gemakkelijk te corrigeren zijn. De filterparameters zijn dan immers al berekend met een foutieve preemfase.

Op grond van deze overweging is dan ook gekozen voor een vaste preemfase van -0.9, zowel voor stemhebbende als voor stemloze frames. Daarnaast blijkt een vaste preemfase ook in de praktijk het beste te voldoen met betrekking tot de analyseresultaten voor de uiting als geheel. We willen graag dat zoveel mogelijk frames precies M/2 toegevoegd complexe nulpunten hebben die als antiresonanties in het resogram terug te vinden zijn. Dan zijn er ook zo weinig mogelijk frames waarin een of meer nulpuntenparen van het analysefilter gebruikt zijn om globale hellingen vlak te maken. In dit opzicht blijkt een vaste u van -0.9 het hoogste aantal frames op te leveren waarvan alle filtersecties toegevoegd complexe nulpunten hebben en daarmee spectrale resonanties neutraliseren.

Een illustratie daarvan is gegeven in fig. 4.11, waarin voor de zin "ieder half uur komt hier 'n bus langs" de resogrammen zijn weergegeven voor verschillende waarden van u. We vinden daarin voor u = +0.9 (onderste resogram) meestal maar vier, soms maar drie van de vijf paren als antiresonanties terug. Voor frekwenties boven 4 kHz, aangegeven met de pijl, is het resogram vrijwel leeg. Als we dit vergelij-

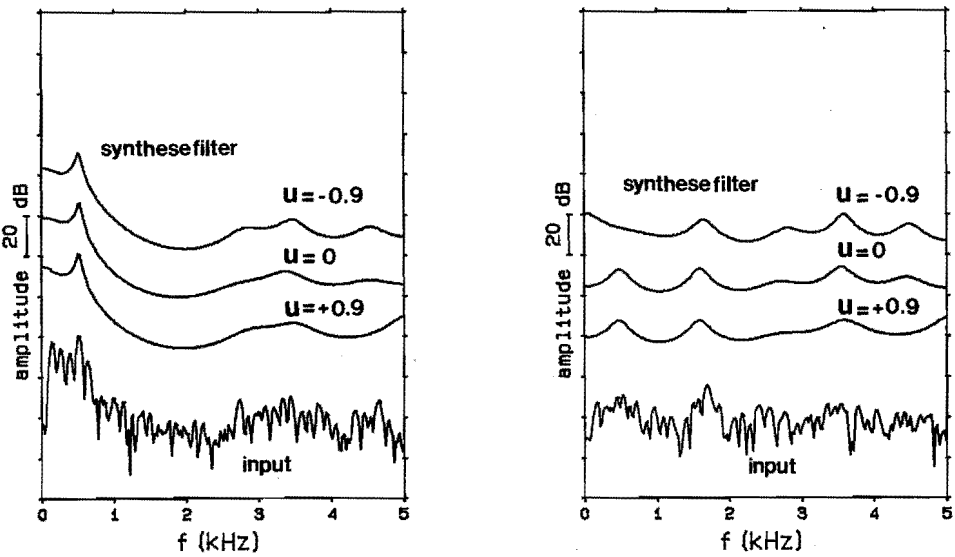

Fig.4.10. Voorbeelden van energiespectra van segmenten uit lopende spraak. Links voor een stemhebbend en rechts voor een stemloos frame. Van boven naar beneden: geinverteerde analysefilters, berekend na pre-emfase van resp. -0.9, 0 en +0.9 (steeds 20 dB verschoven). De onderste curves zijn de inputspectra zelf (40 dB verschoven).

ken met het bovenste resogram, dan blijkt boven 4 kHz voor u = -0.9 globaal steeds een resonantie meer aanwezig te zijn, met uitzondering van een paar stemloze fragmenten. Kennelijk is bij u = +0.9 vaak een 2^e-orde sectie nodig geweest om tussen 4 en 5 kHz een reële (en in dit resogram niet weergegeven) bijdrage te leveren ter neutralisering van de verkeerde pre-emfasehelling. De waarde u = 0 levert, getuige het middelste resogram, resultaten die ruwweg tussen beide vorige inliggen. Hier zien we meer resonanties dan bij u = +0.9, maar toch nog minder dan bij u = -0.9, vooral bij de niet-nasale stemhebbende frames.

Concluderend kunnen we stellen dat een pre-emfasekonstante u van -0.9 voor de overgrote meerderheid van de stemhebbende frames de beste keuze is en het hoogste aantal resonerende deelfilters oplevert. Toepassing van een bivalente u, waarvan de waarde afhangt van de uitkomst van de stem/stemloosdetector, is theoretisch beter maar kan praktische bezwaren opleveren waneer die detector niet feilloos werkt. Daarom is bij de standaardanalyse een vaste waarde voor u van -0.9 gekozen. Uit informele luisterproeven is overigens gebleken dat de keuze van de pre-emfasekonstante weinig kritisch is.

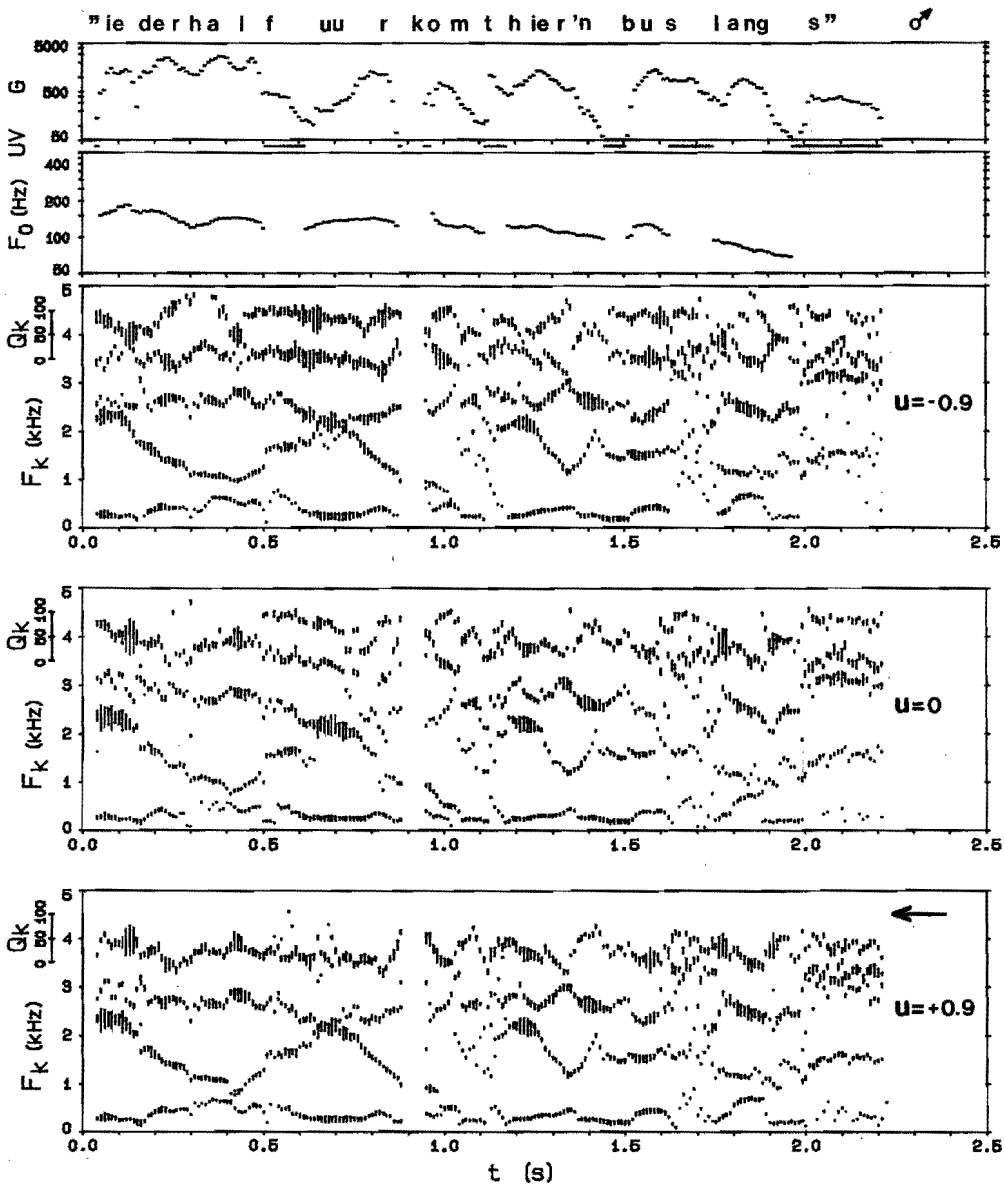

Fig.4.11. Analyseresultaten na pre-emfase van -0.9 (boven), 0 (midden) en +0.9 (onder).

4.3.3. Invloed van het aantal filtercoëffciënten M

Terloops hebben we in hoofdstuk 2 gesteld dat de mate waarin het analysefilter in staat is om het spectrum aan de uitgang vlak te strijken, toeneemt met het aantal coëfficiënten M dat we aan het filter ter beschikking stellen. Een voorbeeld hiervan geeft fig. 4.12a t/m d, waarin een frame uit de m van "komt" is geanalyseerd met resp 6, 10, 18 en 30 coëfficiënten. De bovenste plaatjes geven steeds het energiespectrum van het ingangssignaal, samen met het inverse analysefilter (bovenste gladde curves). In de onderste plaatjes is het energiespectrum van het restsignaal van het betreffende analysefilter weergegeven. Goed is te zien hoe voor toenemende M de omhullende van het restspectrum steeds vlakker wordt. De geresynthetiseerde spraak zal dus, wat z'n energiespectrum betreft, steeds beter overeenkomen met het ingangssignaal. Dat wil echter niet zeggen dat dan ook de analyse met een hoge M zinvol of optimaal is en de vraag is welk aantal coëfficiënten in de praktijk het best voldoet. Het bepalen van veel coëfficiënten vergt alleen maar rekentijd, zonder de geresynthetiseerde spraak wezenlijk te verbeteren. Dat strookt ook niet met ons streven om het spraaksignaal zo zuinig mogelijk te beschrijven; we willen juist de informatiestroom zoveel mogelijk beperken en perceptief onbelangrijke details in het spectrum weglaten. Te weinig coëfficiënten zal ten koste gaan van de kwaliteit van de geresynthetiseerde spraak. Bij de keuze van het optimale aantal zullen luisterproeven dan ook het belangrijkste criterium vormen. Los daarvan zijn er wel een tweetal overwegingen relevant die de keuze van M motiveren.

Allereerst zal het optimale aantal filtercoëfficiënten M afhangen van het totale frekwentiegebied, en dus van de samplefrekwentie f_s waarmee het spraaksignaal is opgenomen in de computer. Een spectrum dat zich uitstrekt tot 10 kHz vergt uiteraard meer parameters dan een telefoonbandbreedte van 3 kHz. De volgende overweging kan dan leiden tot een indicatie van het aantal filterparameters M. In hoofdstuk 2 zagen we dat het analysefilter op een bepaald tijdstip de M voorafgaande samples weegt of convolueert met zijn coëfficiënten. M geeft dus het tijdbereik aan, waarover het filter de samples onthoudt. Markel en Gray (1972) hebben laten zien dat dit bereik minstens 2 keer zo groot moet zijn als de tijd die een geluidsgolf nodig heeft om het akoestisch filter dat het mondkanaal in feite is, te doorlopen. Bij een geluidssnelheid c van 340 m/s en een gemiddelde lengte L van het mondkanaal van ongeveer 0.17 m is die tijd, nodig om het traject van stembanden tot lippen te af te leggen, 0.5 ms. Het filtergeheugen moet zich dus over tenminste 1 ms uitstrekken. Bij een samplefrekwentie f_s van 10 kHz zijn dat 10 samples, bij 6 kHz 6 enz. Dus kunnen we als

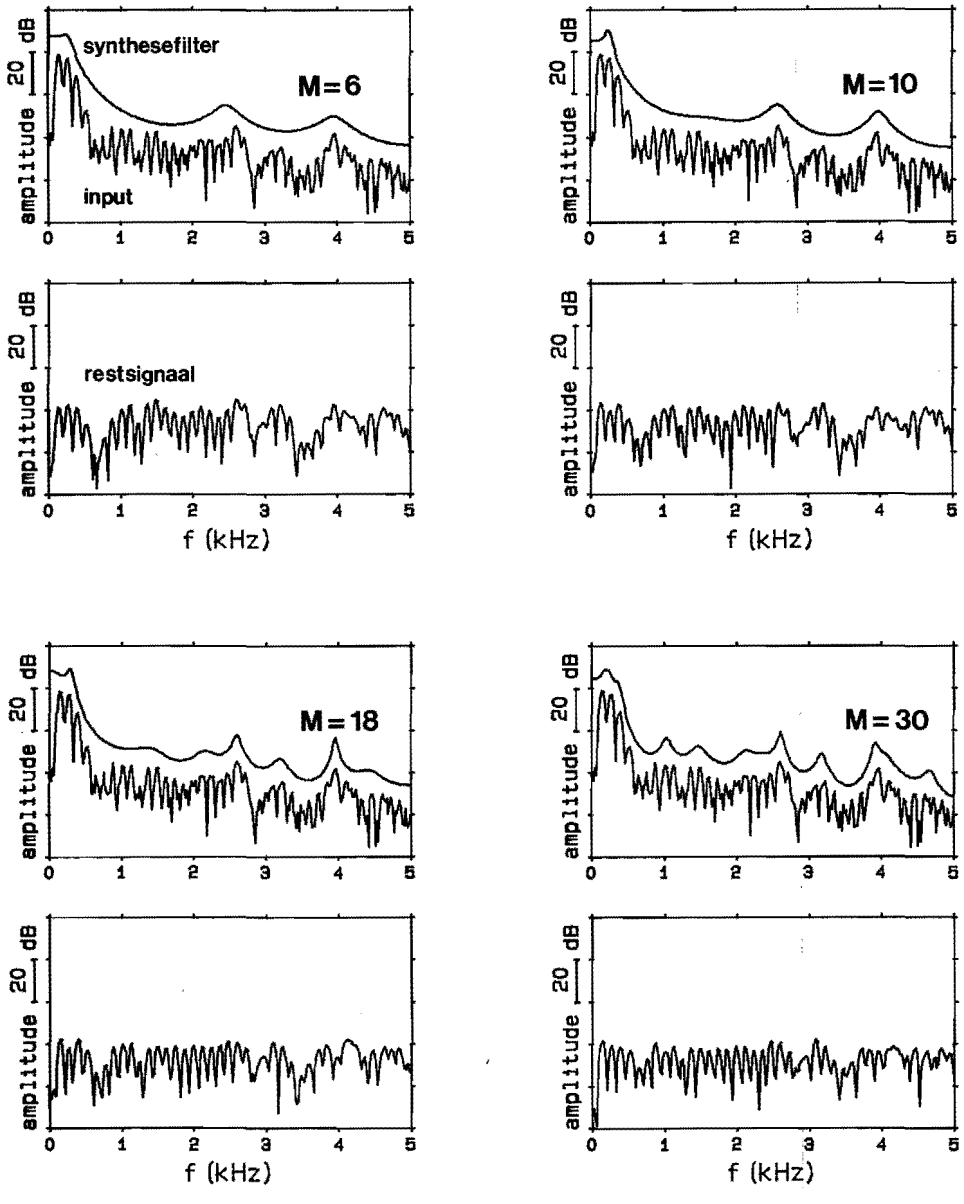

Fig.4.12. Energiespectra van geinverteerde analysefilters (bovenste curves) met daaronder (20 dB verschoven) het inputspectrum voor een fragment uit de m van "komt". Daaronder het spectrum van het restsignaal, na analyse met resp. M = 6, 10 (standaard), 18 en 30 filtercoëfficiënten.

vuistregel hanteren: '10 coëfficiënten per 10 kHz samplefrekwentie', althans voor een mannenstem. Bij vrouwen is het mondkanaal gemiddeld ongeveer 20% korter dan bij mannen. Dus mag het filter ook 20% korter van memorie zijn dan bij mannenspraak en dat levert voor vrouwenspraak dus de vuistregel: '8 coëfficiënten per 10 kHz samplefrekwentie'.

Uit het voorgaande zal duidelijk zijn dat we voor veel stemloze spraaksegmenten met minder parameters zouden kunnen volstaan omdat dan vaak slechts een klein deel van het mondkanaal een rol als akoestisch filter speelt.

Er is nog een tweede overweging waarop we het aantal parameters van het filter kunnen baseren. Het liefst willen we juist voldoende coëfficiënten om, na afsplitsing van de 2^e orde secties, alleen die resonanties weer te geven die als formant in het spectrum terug te vinden zijn. Het is niet zinvol en ook niet nodig om M veel hoger te kiezen dan 2 keer het aantal formanten dat in het gehele frekwentiegebied voor komt. Eén 2^e orde sectie kan immers een formant voor z'n rekening nemen. In de fonetiek worden in gewone stemhebbende mannenspraak tot 5 kHz meestal 5 formanten onderscheiden en bij vrouwenspraak 4. Dat komt dus overeen met eerdergenoemde vuistregel: 5 resp. 4 resonanties per 10 kHz samplefrekwentie.

In fig. 4.13 en 4.14 zien we dat deze vuistregel voor de mannen- en de vrouwenstem aardig bevestigd wordt. Van boven naar beneden zijn weergegeven: de bronparameters (amplitude, stemloos en grondtoonfrekwentie) en de resogrammen met resp. M = 8, 10, 12, en 18. Als we in fig. 4.13 de analyse van de mannenstem met M = 8 vergelijken met die van M = 10 zien we globaal dat er vooral bij de klinkerfragmenten een spoor bijgekomen is. Dat is doorgaans ook vrij breed en betekent dus een relatief scherpe, hoge resonantiepiek. Ook is de traceerbaarheid van de afzonderlijke resonanties bij M = 10 groter dan bij M = 8, doordat gaten in de trajecten worden opgevuld; de sporen vertonen bij M = 10 meer continuiteit. Analyseren met 8 filtercoëfficiënten lijkt dus aan de krappe kant. Verdere verhoging tot M = 12 levert daarentegen niet veel verbetering meer. Als er al een resonantie bij komt, is dat vaak op een enkele punt met een lage kwaliteitsfaktor (smalle stip in het resogram). Ter illustratie is in het onderste resogram nog het resultaat weergegeven van een analyse met M = 18, dus maximaal 9 resonanties. Hier is het weer veel moeilijker om de resonanties te volgen; het plaatje slibt dicht met details. Wanneer we met nog meer coëfficiënten zouden analyseren zouden zelfs de afzonderlijke harmonischen in het spectrum zichtbaar worden.

We zien dus vooral aan de plaatjes voor M = 10 en M = 12 dat er in principe voor de stemhebbende stukken op iedere kHz ruwweg een duidelijke resonantie (formant) voorkomt. Bij de stemloze stukken is dat

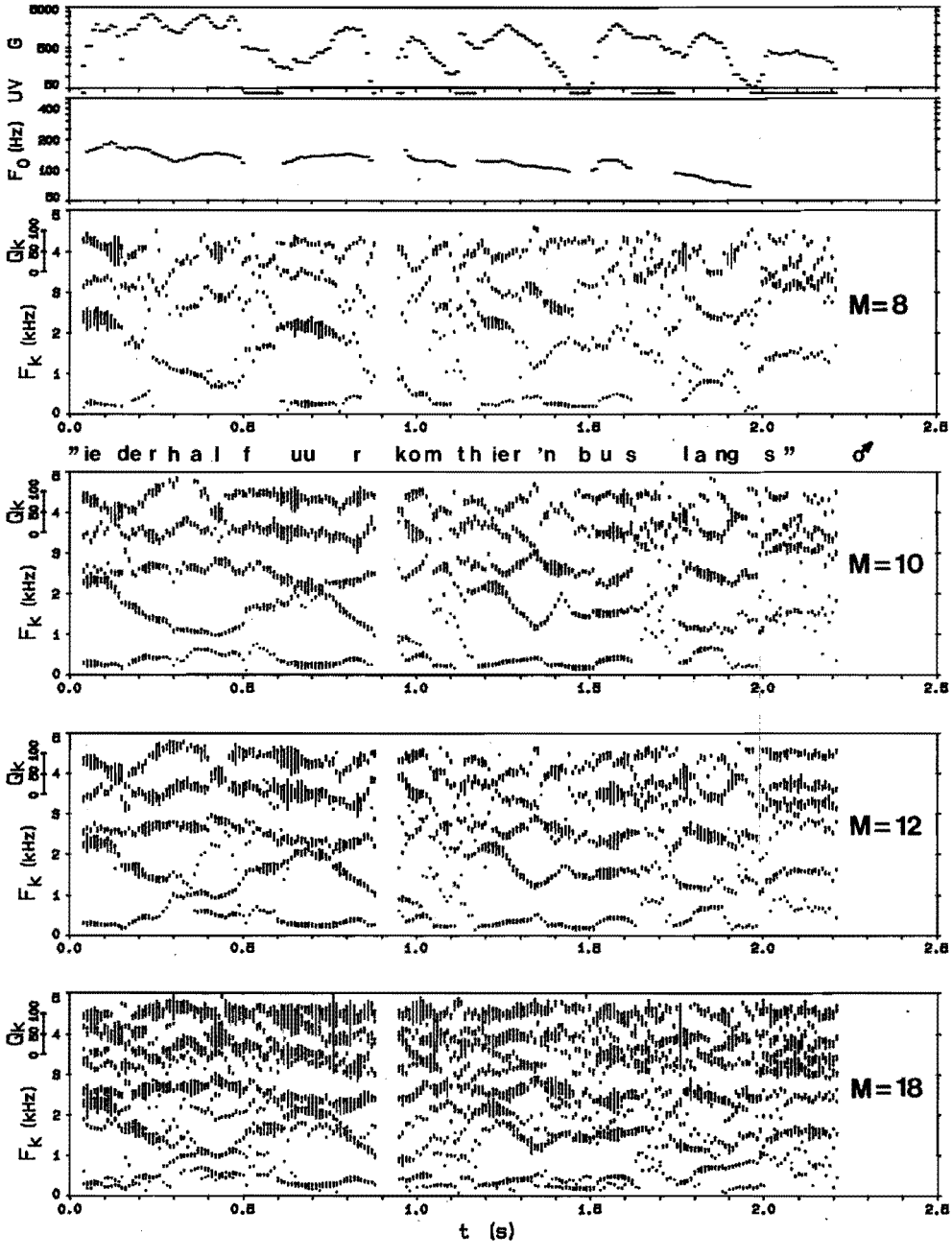

Fig.4.13. Analyseresultaten voor mannenspraak, na analyse met resp. M = 8, 10 (standaard), 12 en 18 filtercoëfficiënten.

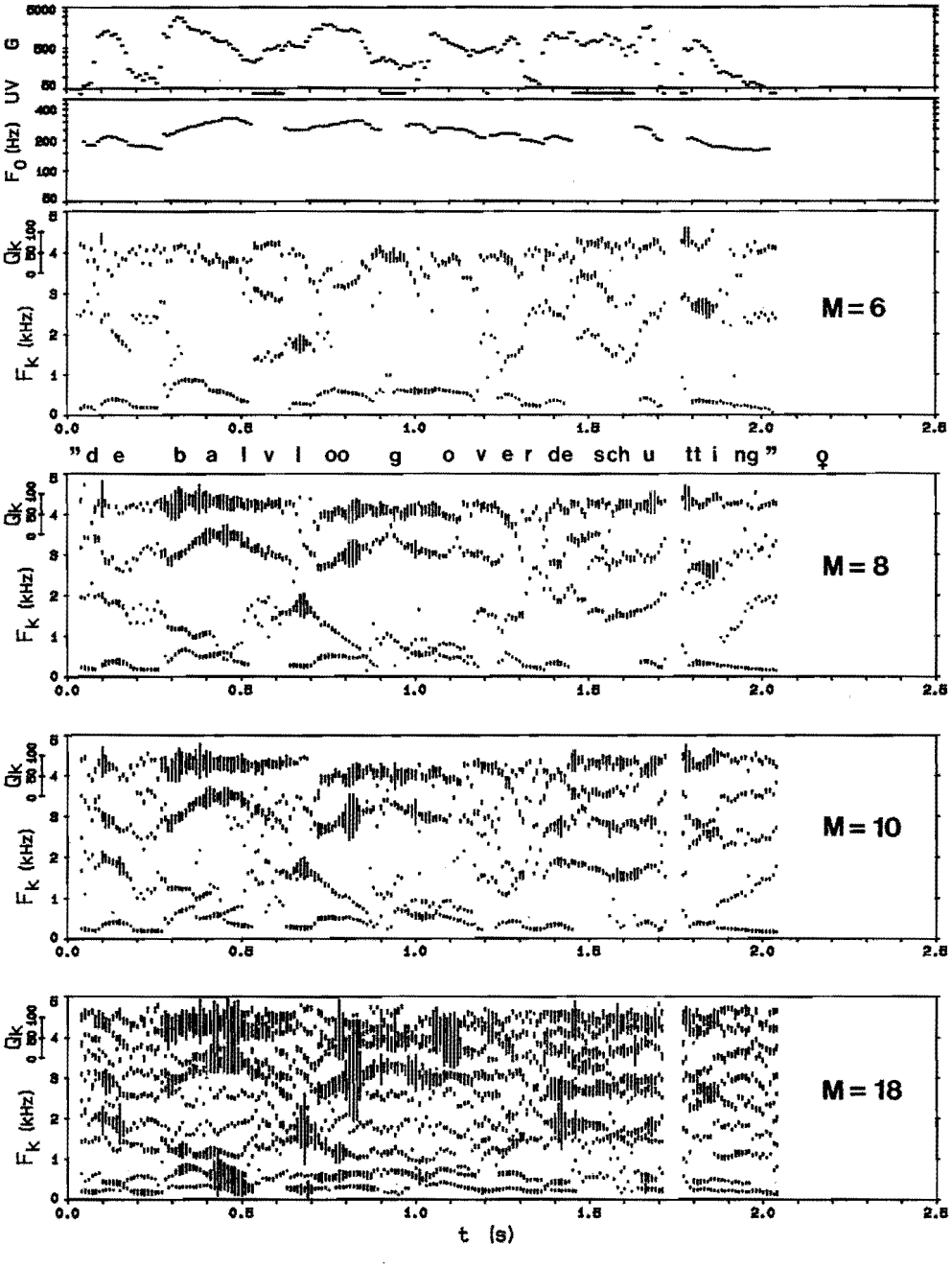

Fig.4.14. Analyseresultaten voor vrouwenspraak, na analyse met resp. M = 6, 8 (standaard), 10 en 18 filtercoëfficiënten.

veel minder het geval.

Voor de vrouwenstem is in fig. 4.14 te zien dat er nu, globaal gesproken, een resonantie minder is. Tussen 0 en 5 kHz zien we meestal niet meer dan 4 resonanties, ook bij M = 10, waarbij er in principe 5 door het filter zouden kunnen worden weergegeven. De vuistregel levert hier dus 8 filtercoëfficiënten bij een samplefrekwentie van 10 kHz.

De conclusie die we op grond van bovengegeven overwegingen trekken is dat voor stemhebbende mannenspraak 10 en voor vrouwenspraak 8 filtercoëfficiënten per 10 kHz samplefrekwentie een goede keus is voor het analysefilter. Voor stemloze stukken spraak zou in principe met minder parameters kunnen worden volstaan. Echter net als bij de keuze van de pre-emfasekonstante zijn er praktische nadelen aan verbonden om tijdens het analyseproces te werken met een variabel aantal coëfficiënten. Dit aantal M zou dan gebaseerd moeten worden op de uitkomst van een niet feilloos werkende stem/stemloosdetector, waardoor correctie achteraf niet meer mogelijk is. Bij onze standaardanalyse zien we daar dan ook van af en werken met een vaste M. Het feit dat daarmee de stemloze stukken wat overbedeeld zijn weegt enigszins op tegen het feit dat die stukken met een niet optimale pre-emfase worden geanalyseerd.

4.3.4. Invloed van de frameperiode T_f

Bij de analyse bepalen we de parameters voor een stukje spraaksignaal ter lengte van N samples in het analysevenster en schuiven dan dit venster op over een stapgrootte T_f, de frameperiode. Deze T_f is binnen zekere grenzen weinig kritisch. In principe hangt T_f enigszins samen met de lengte N van het analysevenster. Het heeft weinig zin om de frameperiode veel korter te kiezen dan overeenkomt met N. In grote delen van het signaal verandert dan zó weinig bij de opeenvolgende stappen dat zo'n korte frameperiode nauwelijks méér details over het parameterverloop oplevert.

In fig. 4.15 is dit geïllustreerd met de analyseresultaten voor het spraakfragment "komt hier 'n bus la(ngs)" met een frameperiode van resp. 4, 10 (standaard) en 40 ms, alle geanalyseerd met de (standaard) vensterlengte van 25 ms. We zien aan de amplitude al hoe zo'n korte frameduur weinig meer is dan een interpolatie tussen de punten die zijn verkregen bij de standaard frameperiode. Datzelfde zien we ook voor de filterparameters in de resogrammen. De korte frameperiode van 4 ms levert nauwelijks meer details. Alleen enkele onrustige stukken, waar resonanties verdwijnen of weer opduiken, zoals bij de overgang o-m en m-t van "komt" levert de korte T_f in het bovenste resogram wat extra resonanties, zij het met een meestal erg lage kwaliteitsfaktor.

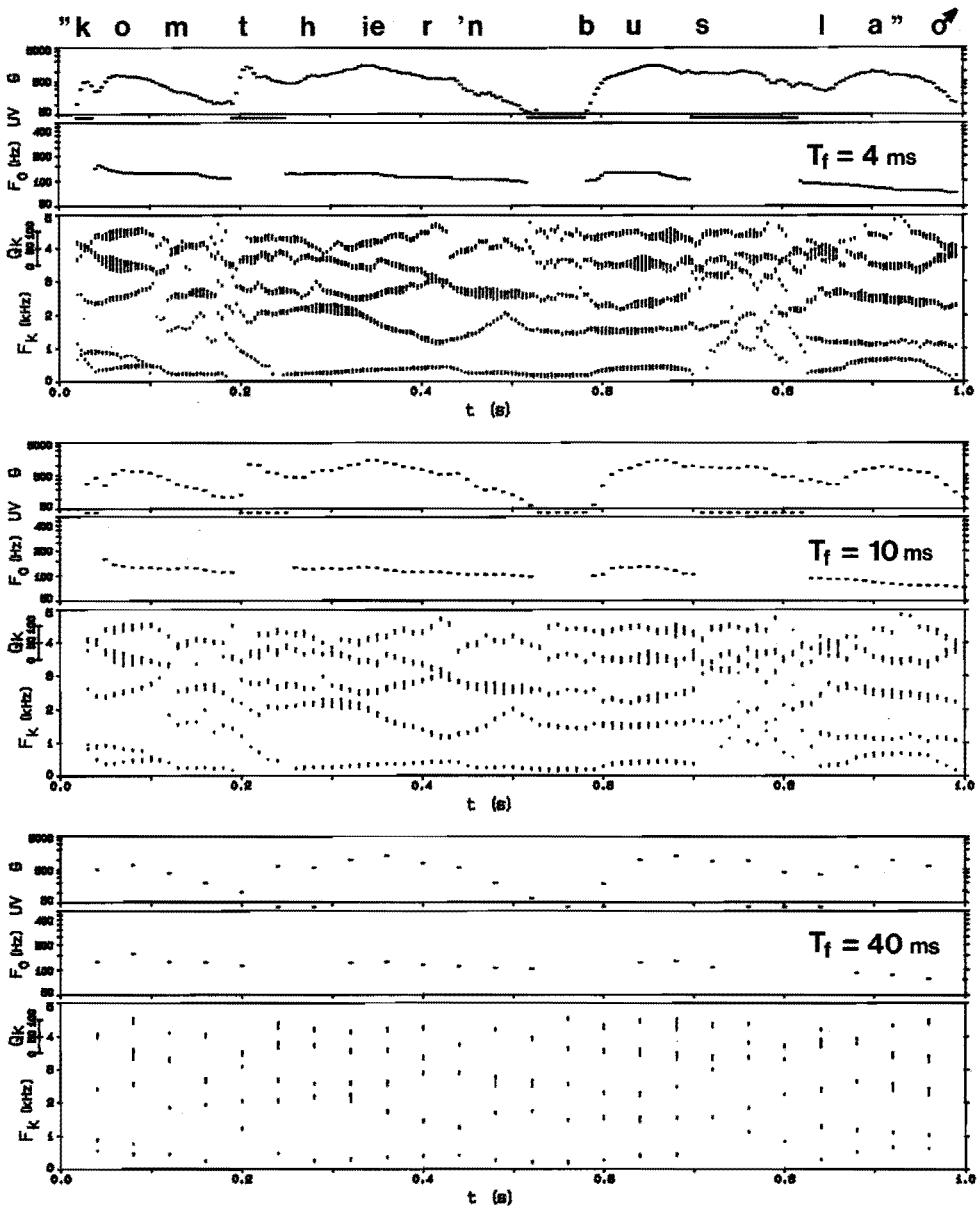

Fig.4.15. Analyseresultaten voor het fragment "komt hier 'n bus la-(ngs)", na analyse met een korte frameperiode van 4 ms (boven), de standaardwaarde van 10 ms (midden) en een lange frameperiode van 40 ms (onder).

Bij de overgang r-'n van "hier 'n", waar de 5 resonanties overgaan in 4 helpt ook een korte T_f niet; er blijft daar een scherpe discontinuiteit in de 5^e en 4^e resonantie.

Uit deze plaatjes blijkt dat korte frameperiodes en daarmee gepaard gaande extra rekentijd de getallenstroom doorgaans alleen maar vergroten, zonder dat daar veel winst aan detailinformatie tegenover staat. Aan de andere kant levert het vergroten van de frameperiode tot bijvoorbeeld 4 keer de standaardwaarde van 10 ms het bezwaar dat dan sommige kortdurende verschijnselen in het ingangssignaal worden gemist. Korte stemloze stukjes van 10 tot 30 ms (van onder meer plofklanken) kunnen bij een frameperiode van 40 ms uiteraard niet meer goed worden gerepresenteerd, zoals uit fig. 4.15 kan worden nagegaan waarin, vergeleken met T_f = 10 ms, drie frames zijn overgeslagen. Steile overgangen worden dan te slap en de geresynthetiseerde spraak klinkt slecht geartikuleerd, enigszins te vergelijken met dronkemansspraak.

Ook nu geldt weer dat ook bij de keuze van de frameperiode luisterproeven de doorslag hebben gegeven. De beperking van de informatiestroom door een grote frameperiode moet worden afgewogen tegen de daarbij optredende mogelijke vermindering van de spraakkwaliteit.

De conclusie die we hier trekken is dat met een standaardlengte van het analysevenster van 25 ms een frameperiode van 10 à 20 ms een geschikte waarde is, waarmee voldoende details kunnen worden weergegeven. Als standaardwaarde gebruiken we in ons systeem 10 ms, die praktische voordelen heeft boven bv 15 of 20 ms en waarmee we ook voor zeer snelle sprekers aan de veilige kant zitten.

We zijn hiermee aan het einde gekomen van de fysische evaluatie van het systeem en gaan over tot de perceptieve evaluatie.

5 PERCEPTIEVE EVALUATIE

In hfdst 0 hebben we ons ten doel gesteld een systeem te ontwikkelen dat synthetische spraak genereert die perceptief niet is te onderscheiden van de oorspronkelijke spraak.

Uit informele luisterproeven is gebleken dat bij lopende spraak die overeenkomst soms zo groot is dat ongetrainde luisteraars geen verschil tussen beide horen. Maar vaak zijn er wel duidelijke verschillen hoorbaar tussen de geresynthetiseerde spraak en de oorspronkelijke, ook voor ongetrainde oren. Hoe groot die verschillen zijn hangt niet alleen af van de opnamecondities (ruis, nagalm, achtergrondlawaai, meerdere sprekers) en de luistercondities (koptelefoon versus luidspreker) maar vooral ook van de spreekstem.

In dit hoofdstuk zullen we het systeem perceptief evalueren door meting van de verstaanbaarheid van de resynthese in vergelijking met de inputspraak van het systeem. Verstaanbaarheid is slechts één deelaspect van de spraakperceptie, dat wel noodzakelijk is maar niet voldoende voor spraakkommunikatie. Ook andere kwaliteitsfactoren zoals natuurlijkheid zullen bijdragen tot de snelheid en het gemak waarmee spraakuitingen verstaan en begrepen worden. Deze factoren zijn echter niet eenvoudig te definiëren of eenduidig te relateren aan fysische eigenschappen van spraakgeluid (Flanagan, 1972).

Diverse methodes om subjektieve spraakkwaliteit te bepalen (voorkeurstest, vergelijkende test, indeling in categorieën, acceptabiliteitstest enz.) laten we hier verder buiten beschouwing; geen van deze tests is algemeen geaccepteerd of aanbevolen voor het evalueren van (synthetische) spraak in de (meest recente) IEEE (1969)-aanbeveling voor het meten van spraakkwalieit. We beperken ons dus tot een evaluatie van ons systeem door middel van verstaanbaarheidsmetingen en zullen in de volgende paragrafen eerst een kort overzicht geven van enkele methodes die hiervoor in de loop der jaren zijn voorgesteld. Daarna recapituleren we de belangrijkste resultaten die één van deze methodes, de spraak-interferentietest van Nakatani en Dukes (1973), toegepast op ons analyse-resynthesesysteem, voor één spreker heeft opgeleverd.

Hoofdthema van dit hoofdstuk vormt een consonant-herkenningstest waarmee voor één spreker is nagegaan in hoeverre afzonderlijke medeklinkers van korte spraakuitingen door het systeem worden aangetast zodat ze niet meer correct worden herkend. Daarbij zullen we ons vooral concentreren op verschillen die diverse versies van analyse en resynthese voor de herkenning van consonanten opleveren. We zullen aantonen dat de fysisch optimale waarden van de vaste modelparameters ook perceptief optimaal zijn. We besluiten het hoofdstuk met een

samenvatting van de belangrijkste resultaten, conclusies en een nabeschouwing.

5.1. SPRAAKVERSTAANBAARHEID

Onder spraakverstaan verstaan we hier het herkennen van spraak of spraakklanken uit spraakgeluid, zodanig dat luisteraars die mondeling of schriftelijk kunnen reproduceren of aanwijzen uit antwoordalternatieven.

In algemene zin kunnen we herkennen opvatten als het aktiveren van een waarnemingsconcept dat door eerdere waarnemingen in het brein is vastgelegd en tengevolge van een stimulus als responsie beschikbaar komt (b.v. Morton 1969, Bouma 1976). Behalve van stimuluseigenschappen is herkenning ook vaak sterk afhankelijk van andere faktoren. Zo kan de context waarin een stimulus eventueel wordt aangeboden herkennen vergemakkelijken. Ook neemt i.h.a. de kans op herkenning toe door herhaalde aanbieding of door verkleinen van het aantal antwoordalternatieven. Verder worden stimuli met een hoge gebruiksfrekwentie in bepaalde gevallen iets gemakkelijker herkend dan minder frekwente.

De lengte van de te herkennen spraak(klanken) kan variëren van losse fonemen tot komplete zinnen en het percentage correct herkende eenheden staat bekend als de articulatiescore. Naarmate de eenheden langer zijn neemt ook het aantal onbekende bijdragen aan de herkenning toe. Onzinsyllaben geven in principe een betere indicatie van de correct herkende fonemen dan woorden of zinnen. Bij deze laatste spelen context, woordfrekwentie, persoonlijke voorkeur e.d. een belangrijke, maar meestal onbekende, rol.

De klassieke methode ter bepaling van spraakverstaanbaarheid maakt gebruik van losse bestaande woorden (Egan 1948), die fonetisch gebalanceerd (PB-woorden) zijn, dwz dat de frekwentie waarmee de klanken in de testwoorden voorkomen overeenkomt met hun gebruiksfrekwentie in spreektaal. Bezwaren die tegen deze test kunnen worden aangevoerd zijn dat eerdergenoemde factoren een onbekende invloed op de articulatiescore hebben en dat de proefpersonen intensief getraind moeten worden om tot betrouwbare uitkomsten te komen.

Om die bezwaren enigszins terug te dringen is door Fairbanks (1958) de zgn rijmtest ontwikkeld, bestaande uit zes lijsten van 50 betekenisdragende CVC woorden, met als C een medeklinker en als V een klinker. Alleen de beginconsonant van de gehoorde woorden moet door de personen worden ingevuld op een antwoordformulier, waarbij ze vrij kunnen kiezen uit alle mogelijke consonanten.

In de zgn gemodificeerde rijmtest (MRT) is deze vrije keuze later door House ea (1965) ingedamd tot een keuze uit zes. Per lijst met bestaande woorden moet van 25 woorden de beginconsonant en van de

andere 25 de eindconsonant worden ingevuld. De zes antwoordalternatieven zijn vervolgens door Griffiths (1967) zó samengesteld dat ze zoveel mogelijk slechts in één onderscheidend fonetisch kenmerk van elkaar verschillen.

Door Voiers (1977) is tenslotte het aantal van zes alternatieve antwoorden verder teruggebracht tot telkens twee. Deze test, de diagnostische rijmtest (DRT), bevat 96 woordparen, 16 voor ieder van de 6 fonetische kenmerken. De proefpersonen krijgen van ieder (visueel aangeboden) paar één woord te horen en moeten dan aangeven welk van beide ze hebben gehoord. Beide (betekenisdragende) woorden verschillen alleen in beginconsonant. Voordeel van rijmtests boven tests met PB-woorden is dat ze met relatief weinig stimulusmateriaal snel en eenvoudig zijn af te nemen en dat ook ongetrainde luisteraars goed reproduceerbare en betrouwbare resultaten opleveren. Vooral de DRT is in de Verenigde Staten veel toegepast bij de evaluatie van zowel spraakkommunicatiekanalen als van synthetische spraak. Een nederlandstalige versie, ontwikkeld door Steeneken (1982) is recent gereedgekomen.

In zowel de PB-woordentest als de rijmtest worden uitsluitend beklemtoonde losse woorden gebruikt en deze tests staan dus nogal ver af van gewone lopende spraak. Tests met komplete zinnen leveren meestal een veel hogere articulatiescore, o.a. omdat de luisteraars gebruik kunnen maken van de samenhang tussen de woorden. Vaak wordt dan ook bij gewone zinnen als testmateriaal al snel het plafond van 100% correcte herkenning bereikt, waardoor het onderscheidingsvermogen voor verschillende spraaksoorten klein is. Dit plafondeffect kan worden vermeden door samen met de testzinnen interfererende spraak (Nakatani en Dukes, 1973), ruisachtige spraak (Kalikow et al, 1977) of spraakachtige ruis (Plomp en Mimpen, 1979) aan te bieden en daardoor de verstaanbaarheid te verlagen. De verhouding van testzinnivo en stoornivo waarbij de herkenningsscore is afgenomen tot 50% staat bekend als spraak-interferentiedrempel SIT (Nakatani en Dukes, 1973), c.q. spraak-receptiedrempel SRT (Plomp, 1979). De wijze waarop deze drempels bepaald worden verschilt nogal van elkaar.

Nakatani en Dukes (1973) bieden de proefpersonen tegelijk met de testzinnen een boekpassage aan, voorgelezen door dezelfde spreekstem. De testzinnen zijn samengesteld uit 5 eenlettergrepige, bestaande woorden (en twee niet gescoorde passende lidwoorden), die in toevalsvolgorde zijn ingevuld. Daardoor hebben ze wel de temporele struktuur, intonatie en klemtoonligging van gewone zinnen maar de woorden zijn onsamenhangend, zodat de luisteraars geen steun hebben van hun betekenis.

Kalikow et al (1977) gebruiken 'babbelruis', dat zijn 12 stemmen bij elkaar gevoegd, om de verstaanbaarheid van gewone zinnen te reduceren en Plomp (1979) past gewone ruis toe, waarvan nivo en spectrale omhullende gelijk zijn gemaakt aan het lange-termijn gemiddelde spectrum van de testzinnen. Door de zinnen qua samenstelling zorgvuldig te selecteren en hun nivo zoveel mogelijk onderling gelijk te maken, bereikt Plomp een hoge betrouwbaarheid, met (voor normaalhorenden) een standaarddeviatie in de 50% drempel (SRT) van slechts 1 dB.

Zowel de SIT- als de SRT-methode zijn in principe toepasbaar om verschillende spraaksoorten te evalueren met betrekking tot hun verstaanbaarheid. De gedachte daarbij is dat goede spraak beter bestand is tegen verstoring door andere spraak of ruis dan spraak van minder goede kwaliteit, en dat dit verschil tot uiting komt in verschillende 50% drempels. Nakatani en Dukes (1973) vinden dat het verschil tussen spraakinterferentiedrempels van hifi-spraak en 'aangetaste' spraak monotoon toeneemt met een, door hen eveneens bepaalde, afnemende (subjectieve) acceptabiliteitsscore. Bij telefoonspraak moeten hun testzinnen 18 dB luider zijn om dezelfde 50% verstaanbaarheid te scoren als bij hifi-spraak.

De spraak-interferentietest

De spraak-interferentiedrempel SIT van Nakatani en Dukes (1973) is door ons destijds gekozen als eerste formele verstaanbaarheidstest voor het analyse-resynthesesysteem (Vogten, 1980) o.a. omdat de nederlandstalige rijmtest en spraak-receptiedrempel-test toen nog niet beschikbaar waren en een test met fonetisch gebalanceerde (PB-) woorden uitvoerig getrainde proefpersonen vereist.

De SIT methode is door ons iets gewijzigd met betrekking tot de interfererende spraak. Bij Nakatani en Dukes (1973) bestaat deze uit een boekpassage, voorgelezen door dezelfde stem die ook de testzinnen uitspreekt. Dat betekent dat woorden van de testzinnen die toevallig samenvallen met pauzes of minder luide passages veel minder beinvloed worden dan andere. Ook kan de boekpassage wisselend en oncontroleerbaar de aandacht van de proefpersoon trekken waardoor testwoorden of hele zinnen gemist kunnen worden. Daarom is in onze uitvoering de gesproken boekpassage eerst ontdaan van pauzes langer dan 50 ms en gecopieerd. Daarna zijn beide kopieën over 5 s in de tijd t.o.v. elkaar verschoven, bij elkaar gevoegd en achterstevoren afgespeeld. Dit interfererende spraakgeluid varieert daarmee minder sterk en leidt de aandacht niet af van de testzinnen. Er zijn 250 verschillende zinnen gegenereerd, alle van dezelfde struktuur: (de draad) (reed) (goed)

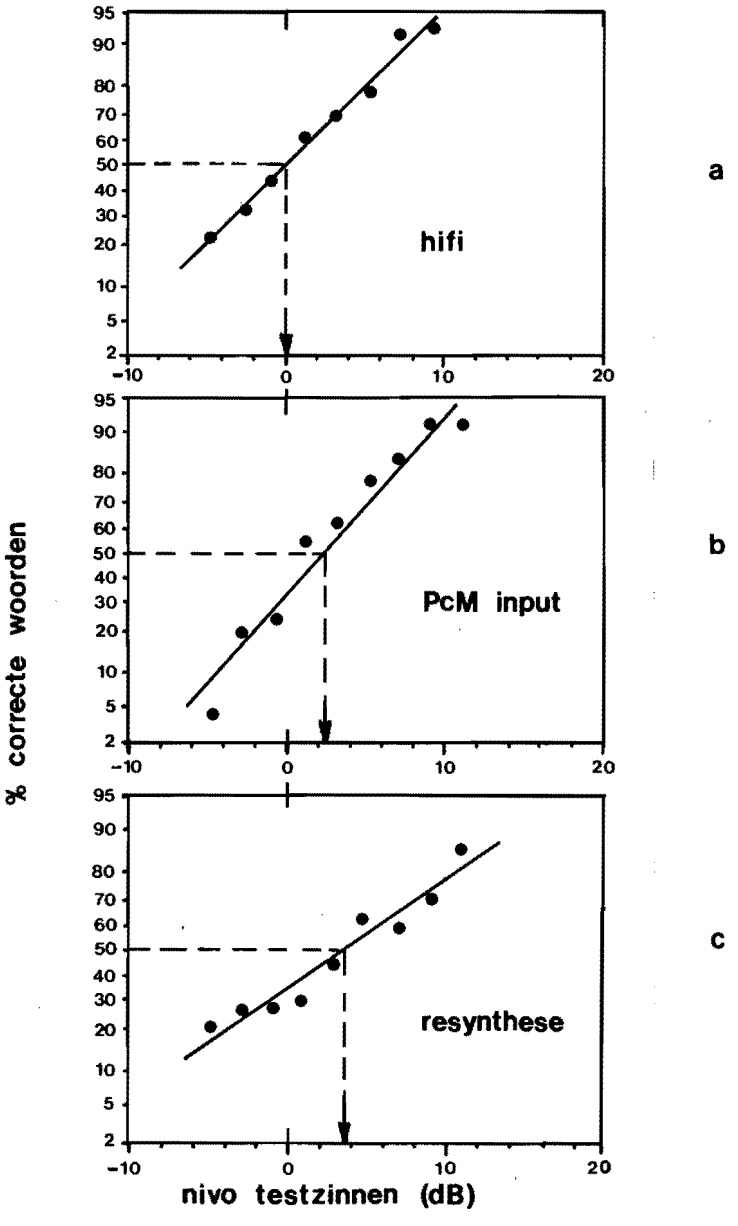

Fig.5.0. Percentage correct herkende woorden als funktie van het nivo waarop ze (in testzinnen) worden aangeboden, bij konstant nivo van het storende spraakgeluid, voor (a): analoge hifi-spraak, (b): PcM gecodeerde spraak (8 kHz, 12 bits, de inputspraak voor het analyse-resynthesesysteem) en (c): geresynthetiseerde spraak.

(voor) (het dal). Werkwoord, zelfstandige naamwoorden en bepalingen zijn vervangen door willekeurige andere eenlettergrepige. Als spraaksoorten zijn o.a. getest: hifi, PcM-gecodeerde (12 bits, 8 kHz samples) en geresynthetiseerde spraak, verkregen door normale analyse (met M = 8 filtercoefficienten) van zowel de testzinnen als het interfererende spraakgeluid. Eventuele toonhoogte- en stem/stemloosfouten zijn niet gecorrigeerd; analyse en resynthese zijn volledig automatisch uitgevoerd. Voor iedere spraaksoort zijn per proefpersoon met 8 à 9 testzinnivo's de articulatiescores voor afzonderlijke woorden bepaald, 10 zinnen (50 woorden) per nivo. Verdere details over procedure, proefpersonen e.d. zijn vermeld in Vogten (1980).

Resultaten

In fig. 5.0 zijn de belangrijkste resultaten uitgezet, gemiddeld over 6 luisteraars. Voor hifi-, PcM- en geresynthetiseerde spraak zijn de correctscores weergegeven als funktie van het nivo van de testzinnen. Het testzinnivo waarbij voor de hifi-versie nog 50% van de woorden herkend werd is als referentie van 0 dB gekozen. We zien dat voor PcM-spraak de 50%-drempel 2.4 dB en voor de resynthese 3.8 dB hoger ligt dan bij de hifi-spraak. Dat betekent dus dat een vermindering in verstaanbaarheid tengevolge van de analyse en resynthese kan worden gekompenseerd door het nivo van de testzinnen t.o.v. het interfererende spraakgeluid met gemiddeld 1.4 dB te verhogen.

De interpretatie van deze uitkomsten is echter niet eenvoudig, juist omdat in deze SIT test interfererende spraak aan de testzinnen wordt toegevoegd. Minder luide delen worden daardoor sterker en vaker gemaskeerd dan luide en zullen dus ook minder bijdragen aan het totstandkomen van de herkenning van de testwoorden. Als het analyse-resynthesesysteem, om welke reden dan ook, spraaksignalen van lage amplitude meer aantast dan delen met hoge amplitude, zou dat nog niet hoeven te leiden tot grote 50%-drempelverschillen tussen input- en outputspraak. Herkenning van de testwoorden gebeurt immers vooral op basis van de niet gemaskeerde fragmenten met relatief hoge amplitude. Deze SIT- (evenals de SRT-) methode is dus niet bij voorbaat de meest geschikte test om eventuele specifieke tekortkomingen van het systeem op te sporen. Wel geeft zij een globale indruk van de vermindering in spraakverstaanbaarheid.

Van de andere genoemde methodes voor het bepalen van de spraakverstaanbaarheid is de MRT in nederlandstalige versie (nog) niet beschikbaar; de DRT inmiddels wel (Steeneken, 1982). Maar het gesloten tweekeuze karakter, de speciale samenstelling van de woordparen en het

feit dat alleen beginconsonanten worden getest, maken de diagnostische waarde van de DRT met betrekking tot specifieke tekortkomingen van ge(re)synthetiseerde spraak nog onduidelijk.

Daarom hebben wij ons, mede gegeven de beschikbare tijd, bij de evaluatie van het analyse-resynthesesysteem gericht op het bepalen van de verstaanbaarheid van afzonderlijke medeklinkers. We gebruiken daarbij een (kwasi) open antwoordkeuze die representatiever is voor het spraakverstaan dan de DRT. In de rest van dit hoofdstuk gaan we daarmee na van welke afzonderlijke medeklinkers de verstaanbaarheid door het analyse-resynthesesysteem wordt aangetast. We beperken ons daarbij niet tot alleen begin- of eindconsonant van beklemtoonde woordjes maar testen ook medeklinkers uit onbeklemtoonde syllaben en overgangen tussen beide.

5.2. CONSONANTHERKENNINGSTEST

5.2.0 Inleiding

In het vorige hoofdstuk hebben we het analyse-resynthese-systeem fysisch geëvalueerd, voornamelijk door aan de hand van resogrammen, golfvormen en energiespectra de resynthese te vergelijken met het oorspronkelijke spraaksignaal. Met deze beelden hebben we geïllustreerd dat door een drietal beperkingen van het analyse-resynthesesysteem, zoals aangegeven in par 4.1, de fysische eigenschappen van een aantal spraakklanken in principe worden aantast. Die klanken zijn met name:

- Plofklanken, met hun relatief snel veranderende eigenschappen, die in principe niet voldoen aan de eis dat het spraaksignaal binnen het analysevenster stationair moet zijn.
- Stemhebbende wrijfklanken, die zowel periodieke als ruisige componenten van dezelfde grootteorde bevatten, terwijl in de resynthese geen kombinatie van periodiek en ruisig brongeluid voorhanden is.
- Nasalen, waarvan het energiespectrum vaak nulpunten bevat terwijl de overdrachtsfunktie van het synthesefilter uitsluitend uit polen bestaat.

We mogen daarom verwachten dat deze klanken na resynthese in principe slechter herkend worden dan de relatief langzaam veranderende klinkers, klinkerachtige medeklinkers en stemloze wrijfklanken, waarvan de fysische eigenschappen niet of weinig door het systeem worden aangetast. Over klinkers kunen we verder kort zijn; het systeem tast deze zó weinig aan dat de herkenning na resynthese niet meetbaar afwijkt van 100%.

Voor de overige klanken, medeklinkers, zullen we in de komende paragrafen nagaan of het systeem aan deze verwachtingen voldoet en welke gevolgen de systeembeperkingen hebben op de perceptie van de geresynthetiseerde spraakklanken. Deze perceptieve evaluatie voeren we uit door de verstaanbaarheid van afzonderlijke medeklinkers in korte stukjes PcM-spraak te vergelijken met verschillende vormen van resynthese. Waar mogelijk zullen we verbanden aangeven tussen de beperkingen van het analyse en synthesesysteem, de daardoor veroorzaakte aantasting van de fysische eigenschappen en de daaruit eventueel voortvloeiende afname in herkenning van medeklinkers.

Concreet gaan we met deze consonantherkenningstest na bij welke medeklinkers de verstaanbaarheid door het systeem wordt aangetast en trachten we uit de eventueel veel gemaakte verwarringen de mogelijke oorzaken voor foute herkenning op te sporen en zo mogelijk aan te geven hoe de geresynthetiseerde spraak eventueel verbeterd zou kunnen worden. Uit de herkenningsscores en de verwarringen die bij incorrecte herkenning van de geresynthestiseerde spraakklanken optreden zullen we

afleiden welke systeembeperkingen hoofdoorzaak zijn voor de gevonden aantasting van consonanten. Deze conclusies zullen worden onderbouwd door naast de normale geresynthetiseerde spraak ook versies te testen waarvan de stem/stemloosparameter achteraf met de hand is gecorrigeerd, om zo de herkenning van een aantal consonanten te verbeteren.

Verder hebben we bij de fysische evaluatie in het vorige hoofdstuk laten zien welke waarden van de 'vaste' modelparameters optimaal zijn en tevens aangetoond dat de procedure om alleen resonerende deelfilters van analyse- en resynthesefilter te verkrijgen weinig of geen invloed heeft op het energiespectrum van de geresynthetiseerde spraak. We mogen dan verwachten dat deze uitkomsten ook perceptief gelden met betrekking tot de consonantherkenning. In dit hoofstuk zullen we laten zien dat de gekozen normale lengte N van analysevenster en frameperiode T_f inderdaad ook perceptief optimaal zijn en tevens aantonen dat de procedure om de filters met louter resonanties te beschrijven perceptief geen nadelige invloed heeft op de herkenning van consonanten.

Van de overige vaste modelparameters zijn in deze tests de preemfase u en het aantal filtercoëfficiënten M niet meer gevarieerd. Informele luisterproeven hebben aangetoond dat variatie van de preemfase geen hoorbare invloed heeft op de geresynthetiseerde spraak. Gegeven de beschikbare tijd is het aantal filtercoëfficiënten beperkt gebleven tot 8 en is deze evaluatie uitgevoerd met slechts één spreker. Hoewel de perceptieve overeenkomst tussen input en resynthese sprekerafhankelijk is, gaat het ons hier in hoofdzaak erom de verschillen tussen de diverse versies van analyse en resynthese vast te stellen. We gaan er dan ook vanuit dat de hierdoor optredende verschillen in consonantherkenning minder sprekerafhankelijk zijn. Hoewel bevestigd door informele luisterproeven, zou dat in formele experimenten nog verder aangetoond kunnen worden.

5.2.1. Methode

De verstaanbaarheid van medeklinkers is bepaald door luisteraars korte uitingen aan te bieden, waarin één medeklinker voorkomt en ze te laten opschrijven welke medeklinker ze hebben gehoord.

Getest zijn de plofklanken p, t, k, b, d; de wrijfklanken g(=ch), f, v, s, z; de nasalen m, n, ng en de approximanten r, l, w, j en h. Elke van deze 18 nederlandse medeklinkers is gekombineerd met 4 beklemtoonde klinkers: lange aa, ie, oe en korte u, (overeenkomend met de hoekpunten en het midden van de klinkerdriehoek) en met de onbeklemtoonde 'neutrale' klinker e (Nooteboom en Cohen, 1976).

Stimulusmateriaal

De medeklinkers zijn aangeboden in 5 verschillende posities: VC, CV, eCV, VCe en Ce, waarin C de betreffende medeklinker is, V een van de 4 klinkers en e de onbeklemtoonde 'neutrale' klinker.
Niet alle medeklinkers zijn in alle posities aangeboden. Op grond van regels voor uitspraak of voorkomen in het nederlands zouden b d v z en h in VC positie, ng in CV en eCV positie en h in VCe positie kunnen worden uitgesloten. De hier gebruikte stimuli staan echter ver van gewone (spreek)taal af. Verder is spreker noch luisteraars gevraagd om zich aan uitspraakregels te houden bij het uitspreken resp. beoordelen van de stimuli. Daarnaast verwachten we, zoals in par. 5.2.0 is uiteengezet, dat juist van laatstgenoemde medeklinkers de herkenning door het systeem wordt aangetast. Daarom is, enigszins willekeurig, besloten een aantal van deze 'onnatuurlijke' kombinaties wèl op te nemen in het stimulusmateriaal. De komplete lijst van aangeboden stimuli is weergegeven in tab. 5.1.

De stimuli zijn door één mannelijke spreker op magneetband ingesproken. Om de vereiste beklemtoonde en onbeklemtoonde fragmenten te verkrijgen zijn de klinker/medeklinkerkombinaties bij het uitspreken verlengd tot "CVt", "tVC", "tVCe", "teCV" en "Cetaat", met als C een van genoemde 18 medeklinkers en als V een van de 4 klinkers.

Na laagdoorlaatfilteren op 4 kHz zijn deze verlengde stimuli met 8 kHz samplefrekwentie gedigitaliseerd (PulscodeModulatie, PcM) in 12 bits per sample en ingelezen in het schijvengeheugen van de rekenmachine. Vervolgens zijn de aanvullingen "t", "te" of "etaat" verwijderd met een interaktief segmenteerprogramma (via auditieve en visuele inspectie op nuldoorgangen in de golfvorm van het signaal) en de zo verkregen stimuli opgeslagen in nieuwe files. Deze vormden de input (versie PcM) voor het analyse-resynthesesysteem.

Spraakversies

Behalve de inputversie PcM zijn nog 9 versies geresynthetiseerde spraak aangeboden.

Eerst is de PcM versie geanalyseerd in a-parameters, zoals beschreven in hoofdstuk 2, met als vaste modelparameters de bij 8 kHz normale waarden: M = 8 coefficienten van het analysefilter, T_f = 8 ms frameperiode (dus niet de standaardwaarde van 10 ms i.v.m. uitgifte via de spraakchip, daarover meer in hfdst. 6), L_w = 25 ms duur van het analysevenster en pre-emfase u = -0.9. Resynthese hiervan leverde de normale A-versie NoA op.

Vervolgens zijn de a-parameters omgerekend naar 2e orde pq-parameters, omgezet in 4 resonerende paren en geordend naar afstemfrekwen-

1e zitting:

CV				Ce	VC			
paa	pie	poe	pu	pe	aap	iep	oep	up
taa	tie	toe	tu	te	aat	iet	oet	ut
kaa	kie	koe	ku	ke	aak	iek	oek	uk
baa	bie	boe	bu	be	aab	ieb	oeb	ub
daa	die	doe	du	de	aad	ied	oed	ud
gaa	gie	goe	gu	ge	aag	ieg	oeg	ug
faa	fie	foe	fu	fe	aaf	ief	oef	uf
vaa	vie	voe	vu	ve	--	--	--	--
saa	sie	soe	su	se	aas	ies	oes	us
zaa	zie	zoe	zu	ze	--	--	--	--
maa	mie	moe	mu	me	aam	iem	oem	um
naa	nie	noe	nu	ne	aan	ien	oen	un
--	--	--	--	--	aang	ieng	oeng	ung
raa	rie	roe	ru	re	aar	ier	oer	ur
laa	lie	loe	lu	le	aal	iel	oel	ul
waa	wie	woe	wu	we	aaw	iew	oew	uw
jaa	jie	joe	ju	je	aaj	iej	oej	uj
haa	hie	hoe	hu	he	--	--	--	--

2e zitting:

VCe				eCV			
iepe	aape	oepe	uppe	epie	epaa	epoe	epu
iete	aate	oete	utte	etie	etaa	etoe	etu
ieke	aake	oeke	ukke	ekie	ekaa	ekoe	eku
iebe	aabe	oebe	ubbe	ebie	ebaa	eboe	ebu
iede	aade	oede	udde	edie	edaa	edoe	edu
iege	aage	oege	ugge	egie	egaa	egoe	egu
iefe	aafe	oefe	uffe	efie	efaa	efoe	efu
ieve	aave	oeve	uvve	evie	evaa	evoe	evu
iese	aase	oese	usse	esie	esaa	esoe	esu
ieze	aaze	oeze	uzze	ezie	ezaa	ezoe	ezu
ieme	aame	oeme	umme	emie	emaa	emoe	emu
iene	aane	oene	unne	enie	enaa	enoe	enu
ienge	aange	oenge	unge	engie	engaa	engoe	engu
iere	aare	oere	urre	erie	eraa	eroe	eru
iele	aale	oele	ulle	elie	elaa	eloe	elu
iewe	aawe	oewe	uwwe	ewie	ewaa	ewoe	ewu
ieje	aaje	oeje	ujje	ejie	ejaa	ejoe	eju
--	--	--	--	ehie	ehaa	ehoe	ehu

Tab.5.1. Lijst van alle aangeboden stimuli. De onderstreepte stimuli vormen bestaande nederlandse woordjes die in de spreektaal een frekwentie van 10 of meer hebben volgens Uit den Bogaart (1975).

tie conform paragraaf 2.5. Dit leverde na resynthese de normale P-versie NoP.

Dan zijn vier A-versies gevormd door analyse met resp. een smal analysevenster L_w = 10 ms (versie SmA), een breed venster L_w = 60 ms (versie BrA), een korte frameduur T_f = 4 ms (versie KoA) en een lange frameduur T_f = 40 ms (versie LaA).

Vervolgens is van beide versies NoA en NoP de stem/stemloosparameter met 'oor en hand' via een interaktief programma gecorrigeerd, zodanig dat een aantal consonanten beter op die van de inputversie leken. Dit leverde resp. de versies CoA en CoP op.

Tenslotte is de normale p-versie NoP nogmaals aan de proefpersonen aangeboden om enige indicatie over leer- of gewenningseffekten te verkrijgen. Deze herhaling zal hier verder worden aangeduid met versie NoP".

Bij alle geresynthetiseerde versies is uit praktische overwegingen steeds een konstante toonhoogte toegepast van 100 Hz. Daarmee is eventuele herkenning van bepaalde combinaties op grond van eventuele (kleine) toonhoogteverschillen uitgesloten.

De verschillende versies zijn na resynthese in files op het schijvengeheugen opgeslagen en vervolgens in toevalsvolgorde, na 12-bits digitaal-analoogomzetting en laagdoorlaatfiltering tot 4 kHz, op magneetband opgenomen.

Procedure

Gezeten in een geluidsarme box beluisterden de proefpersonen de stimuli via bandrecorder (Revox B77) en koptelefoon (Sennheiser HD424), op beide oren hetzelfde signaal en op een komfortabel geluidsnivo (70 dBSL voor de auteur als proefpersoon). Alle luisteraars kregen de spraakversies in dezelfde volgorde aangeboden. Eerst de PcM (inputspraak) en daarna de 9 geresynthetiseerde versies, in de volgorde zoals eerder beschreven: NoA, NoP, SmA, BrA, KoA, LaA, CoA, CoP en NoP". Elke versie nam ruim een kwartier in beslag en was opgesplitst in twee delen: eerst de 145 VC, CV en Ce kombinaties in blokjes van 10 en toevalsvolgorde. Dan, indien gewenst, een korte pauze en vervolgens op dezelfde wijze de 140 VCe en eCV kombinaties. Per 3 seconden werd één stimuluswoordje aangeboden en in de tussenliggende tijd moest de daarin gehoorde consonant op een formulier worden opgeschreven door invullen van: p, t, k, b, d, g of ch, f, v, s, z, m, n, ng, r, l, w, j, of h. De proefpersonen werd verzocht om altijd een antwoord in te vullen. Ieder blok van 10 stimuli werd voorafgegaan door een attentiesignaal en tussen de opeenvolgende versies lag een tijdsbestek van tenminste een uur. Uit de resultaten van de als eerste geteste PcM versie bleek dat de proefpersonen geen enkele moeite hadden met deze

taak. Voor de versies met geresynthetiseerde spraak zijn dan ook geen oefenseries aangeboden.

Proefpersonen

Behalve de auteur namen nog 8 luisteraars deel aan de experimenten, 3 vrouwen en 5 mannen, allen IPO medewerkers met normale gehoorfunktie en in leeftijd variërend van 26 tot 39 jaar. Hiervan hadden er 3 zelden of nooit eerder synthetische spraak beluisterd, de overigen wel, variërend van weinig tot vrij veel. Ook de mate van ervaring met het deelnemen aan dit type experimenten liep sterk uiteen.

5.2.2. Resultaten en discussie

In deze paragraaf gaan we na hoe de herkenning van de afzonderlijke consonanten, gemiddeld over de 8 proefpersonen, door de verschillende versies van analyse en resynthese worden beinvloed. De resultaten van de auteur als proefpersoon zijn hierbij verder buiten beschouwing gelaten.

Eerst vergelijken we de consonantverstaanbaarheid van de (automatisch verkregen) normale resynthese (versie NoP) met de inputspraak (versie PcM). Daarbij zal blijken dat relatief veel fouten zijn gescoord bij die consonanten waarvoor indeling in stemhebbend of stemloos niet goed mogelijk is, omdat ze zowel periodieke als ruisige componenten bevatten van gelijke grootteorde. Dat vooral stem/stemloos-fouten de oorzaak zijn van veel verwarringen zal blijken uit het feit dat de herkenning van een aantal consonanten aanzienlijk verbeterd kan worden door de (automatisch verkregen) stem/stemloosbeslissing met 'oor en hand' te corrigeren. Dat zullen we laten zien door versie CoA en CoP te vergelijken met NoA en NoP.

Vervolgens tonen we aan dat de normaal toegepaste waarden voor analysevenster en frameduur optimaal zijn voor consonantherkenning, door de versies SmA, BrA, KoA en LaA te vergelijken met de normale versie NoP.

Tenslotte zullen we, door de versies NoA en CoA te vergelijken met versie NoP en CoP, laten zien dat het louter resonerend maken van de deelfilters in analyse- en synthesefilter geen nadelige invloed heeft op de verstaanbaarheid van de geteste medeklinkers.

1. De invloed van analyse en resynthese

In fig 5.1 zijn de foutscores uitgezet voor de afzonderlijke consonanten voor de PcM (inputversie), de normale resynthese NoP (output)

en de herhalingsversie NoP". Per consonant is gemiddeld over alle klinkers/posities en over de 8 proefpersonen. De consonanten zijn geordend naar toenemende foutscore voor versie NoP. Uiterst rechts in de plaatjes is het gemiddelde over alle consonanten van de betreffende versie weergegeven. De verticale strepen hebben een totale lengte van 2 keer de standaarddeviatie tussen de gemiddelden van de afzonderlijke luisteraars.

Voor de drie versies zijn in fig 5.2. de verwarringsmatrices weergegeven. De rijen met de aangeboden consonant zijn onderverdeeld in de categorieën plofklanken, wrijfklanken, nasalen en approximanten. In de kolommen staan de geantwoorde medeklinkers in procenten van het totale aantal aanbiedingen. Per medeklinker is ook hier gemiddeld over alle posities/klinkers en proefpersonen. Per rij kan het totaal iets afwijken van 100% omdat is afgerond op hele procenten (0.5 en hoger naar boven).

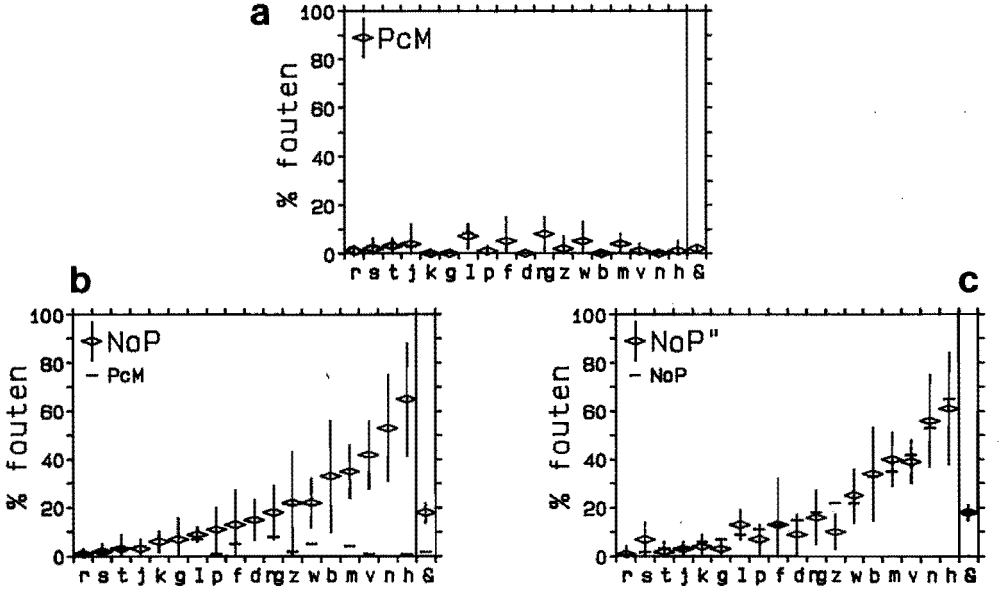

Fig.5.1. Percentage fouten per consonant, gemiddeld over alle klinkers, posities en 8 proefpersonen. De consonanten zijn geordend naar NoP scores. Bij (a) de input PcM-versie, bij (b) de normale resynthese NoP samen met de PcM scores en bij (c) de herhalingsversie NoP" samen met de NoP scores. Uiterst rechts is in ieder plaatje bij & de gemiddelde score over alle consonanten aangegeven. De verticale strepen hebben een lengte van 2 maal de standaarddeviatie tussen de gemiddelden van de afzonderlijke proefpersonen.

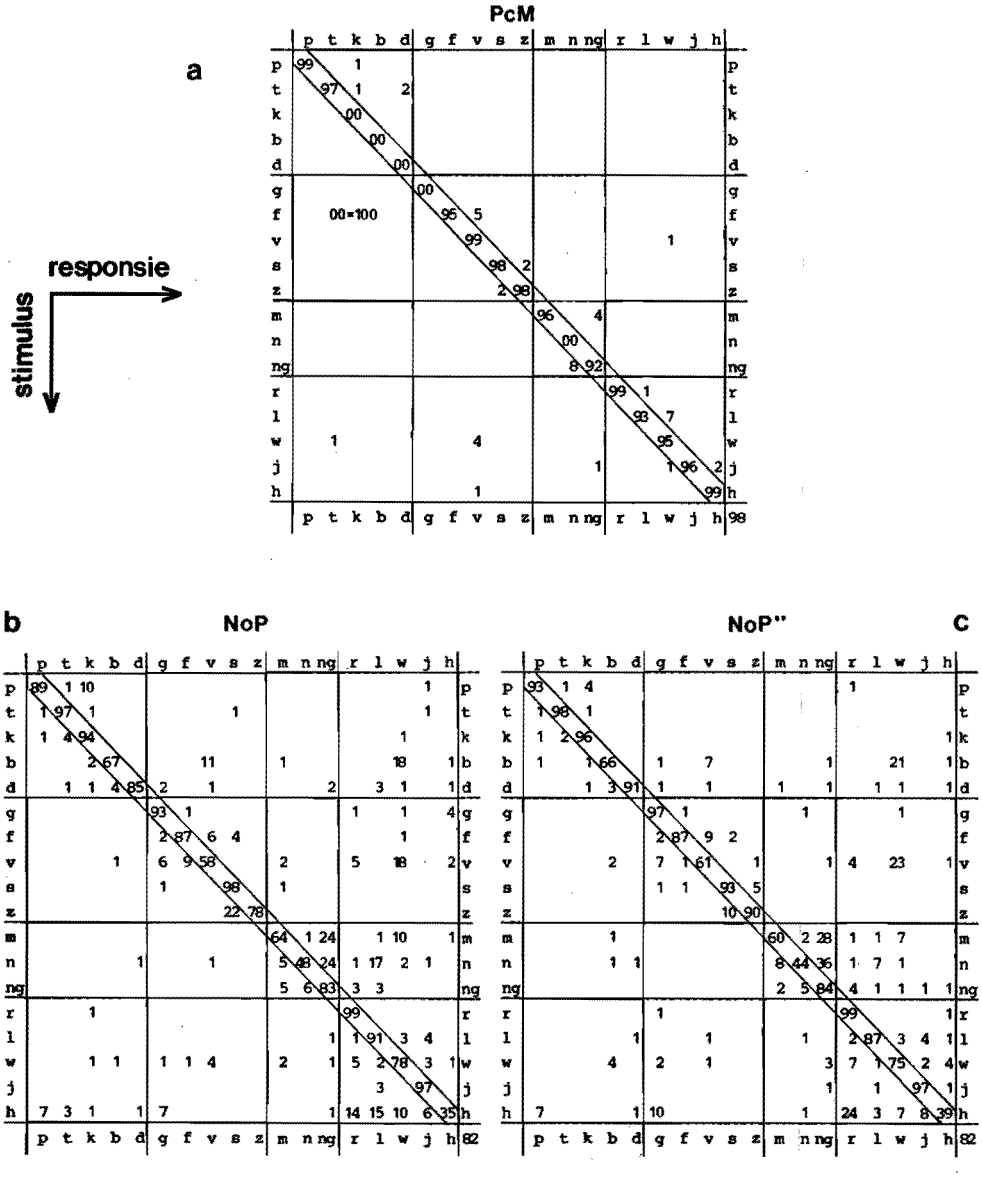

Fig.5.2. Verwarringsmatrices, (a) voor de versie PcM, (b) voor versie NoP en (c) voor de herhalingsversie NoP". Vertikaal de stimuli, horizontaal de responsies, in afgeronde procenten van het totaal aantal aanbiedingen en gemiddeld over alle klinkers/posities en 8 proefpersonen. Rechts onder in iedere matrix het percentage correct, gemiddeld over alle consonanten.

We zien dat de PcM versie weinig problemen geeft; gemiddeld over alle consonanten 98% correct. De f→ v fouten zijn afkomstig van één proefpersoon, evenals de w→ v fouten. Daarentegen liggen de ng→ n en l→ w verwarringen verspreid over vrijwel alle proefpersonen. Verder merken we op dat in de 'onnatuurlijke' VC positie b en d foutloos worden herkend; alleen de ng levert in eCV positie wat meer fouten dan in de overige posities.

De normale resynthese NoP leidt, gemiddeld over alle consonanten, tot 82% correcte herkenning, dus 16% meer fouten dan de PcM versie. Slechter dan gemiddeld scoren de stemhebbende plofklank b, de stemhebbende wrijfklanken v en z, de nasalen m en n en de approximanten w en h. Met welke andere consonanten ze veel verward worden kunnen we nagaan aan de verwarringsmatrix van versie NoP. De stemhebbende plofklank b levert vaak b→ w en b→ v verwisselingen op. In de categorie stemhebbende wrijfklanken wordt de z uitsluitend verwisseld met de s en levert veel z→ s verwisselingen. De v wordt daarentegen veel met andere medeklinkers verward: zowel v→ w als v→ f, v→ g en v→ r komen vaak voor. Bij de nasalen vinden we zowel fouten binnen de eigen categorie, met veel m→ ng, en n→ ng verwarringen, als daarbuiten, met voornamelijk m→ w en n→ l verwisselingen. Bij de approximanten tenslotte komt het antwoord op de w sterk verspreid terecht en ook met de h weten de proefpersonen niet goed raad; naast veel foute antwoorden binnen de eigen categorie zijn er vooral h→ p en h→ g verwarringen.

Tussen de diverse consonantposities hebben we, gemiddeld over alle medeklinkers en de 4 klinkers, geen grote scoreverschillen gevonden; de foutenpercentages variëren bij versie NoP tussen 14% en 21%. Wel treden er voor de verschillende posities grote verschillen op bij de afzonderlijke klinkers. Zo levert de klinker aa in VC-positie gemiddeld slechts 5% foute consonanten op, maar leidt de klinker ie in de VC-positie tot 38% fouten.

Verder vinden we, net als bij de PcM versie, ook in versie NoP geen duidelijke verschillen bij de consonanten b en d tussen de onnatuurlijke VC-positie en de overige posities, maar levert de ng in de tussenpostie eCV wel meer fouten dan in de VCe-positie.

Wanneer we de scores van de herhalingsversie NoP" vergelijken met die van NoP versie NoP (fig 5.2b en c) zien we dat de gemiddelde score voor alle medeklinkers in beide versies gelijk is (18% fouten). Voor de afzonderlijke consonanten is alleen bij de z de foutscore duidelijk lager en de standaarddeviatie kleiner geworden. Deze verbetering blijkt hoofdzakelijk afkomstig te zijn van proefpersoon nr 5, die in versie NoP" 50% minder fouten scoort dan in versie NoP.

Hoewel tussen de versies NoP en NoP", blijkens de verwarringsmatrices in fig 5.2b en c, voor de individuele consonanten soms wel ver-

schillen zijn, zowel in de correctscore per consonant als in de verdeling van de foute antwoorden over de consonanten, is het globale beeld voor beide hetzelfde. Hieruit concluderen we dat binnen de termijn waarin beide versies geresynthetiseerde spraak zijn aangeboden, weinig leer- of gewenningseffekten optreden.

Omdat de verschillen tussen beide versies NoP en NoP" klein zijn zullen we in het vervolg bij de vergelijkingen tussen de verschillende analysemethodes steeds de normale versie NoP als referentie nemen, tenzij expliciet anders vermeld.

Discussie

Verwisselingen die bij incorrecte herkenning optreden hebben sinds Miller en Nicely (1955) een belangrijke rol gespeeld bij het opsporen van kenmerken voor de herkenning van fonemen. De verwisselingspatronen in deze matrices worden verondersteld de perceptieve gelijkenis tussen stimuli te weerspiegelen. Daarbij wordt dan vaak ruis aan de stimuli toegevoegd om voldoende verwarringen te verkrijgen teneinde statistisch betrouwbare conclusies te kunnen trekken. Hoewel de term perceptieve gelijkenis een zekere symmetrie suggereert, blijken verwarringsmatrices lang niet altijd symmetrisch te zijn; stimulus x wordt soms veel vaker als y gerespondeerd dan stimulus y het antwoord x oplevert. Deze asymmetrieën, vaak aangeduid als 'bias', kunnnen niet alleen toegeschreven worden aan fysische of andere eigenschappen (zoals bv. gebruiksfrekwentie) van de stimulus zelf, maar moeten vaak ook gezocht worden in het antwoordproces van de proefpersoon, o.m. door effekten van herhaling en persoonlijke voorkeuren.

Ook in onze verwarringsmatrices vinden we duidelijke asymmetrieën. Ofschoon de verwarringen hier niet kunstmatig vergroot zijn door toevoeging van ruis kunnen we ons afvragen welke van de hierboven en in par. 5.1 genoemde faktoren in zowel stimulus als antwoordproces tot de herkenning c.q. verwisselingen hebben bijgedragen. We zullen die faktoren hier nu achtereenvolgens bespreken.

De invloed van context en herhaalde aanbieding kunnen we buiten beschouwing laten omdat iedere stimulus zonder context en per spraakversie slechts een keer per proefpersoon is aangeboden. Ook de invloed van het beperkte aantal van 18 antwoordalternatieven is hier verwaarloosbaar; proefpersonen hebben door te raden slechts ongeveer 6% kans op een goed antwoord zonder dat er sprake is van herkenning.

Voor persoonlijke voorkeuren zijn in sommige gevallen wel aanwijzingen gevonden, nl. daar waar bepaalde verwarringen grotendeels afkomstig zijn van slechts een proefpersoon, zoals in de PcM-versie de f→g en w→v fouten. We kunnen hierover echter verder weinig concrete uitspraken doen; daarvoor is het aantal verwarringen per proefpersoon

per stimulus te gering. Verder zijn bij de geresynthetiseerde spraak geen grote verschillen tussen de verschillende proefpersonen gevonden.

Rest ons nog na te gaan welke invloed de faktor frekwentie van voorkomen kan hebben gehad op de resultaten. De proefpersonen antwoordden door een van de 18 medeklinkers op te schrijven en het is dus in principe mogelijk dat hoogfrekwente medeklinkers vaker als fout geantwoorde medeklinker voorkomen dan laagfrekwente. We hebben dat nagegaan door voor versie NoP en NoP" van alle consonanten de foute antwoorden (kolomtotalen exclusief diagonaal) te delen door het totaal aantal gemaakte fouten en deze percentages uit te zetten tegen de relatieve gebruiksfrekwentie in de spreektaal volgens Eggermont (1956). Tussen beide blijkt geen verband te bestaan.

Tenslotte is er dan nog de mogelijkheid dat stimuli die bestaande nederlandse syllaben of woorden vormen met een hoge gebruiksfrekwentie beter herkend worden dan laagfrekwente of niet bestaande. Voor gesproken syllaben zijn geen frekwentietellingen beschikbaar, wel voor woorden. In de lijst van stimuli (tabel 5.1, par 5.2.1.1) is aangegeven welke van de 145 CV-, VC- en Ce-stimuli woordjes vormen met een frekwentie van 10 of meer in de spreektaal (Uit den Bogaart, 1975). We kunnen nu nagaan of deze stimuli ook duidelijk minder fouten hebben opgeleverd dan de overige. Ook dat blijkt in het algemeen niet het geval te zijn; noch voor de inputversie PcM, noch voor de beide geresynthetiseerde versies NoP en NoP". Hoewel 21 van de 24 hoogfrekwente stimuli in de groepen CV en Ce vallen is, over alle consonanten gemiddeld, het foutenpercentage niet lager dan voor de VC-groep waarvan slechts 3 stimuli hoogfrekwent zijn. Wel vinden we voor de afzonderlijke consonanten in één geval (de n) dat in de versies NoP en NoP" de bestaande woordjes "na(a)" en "aan" duidelijk minder fouten scoren dan de overige stimuli met de consonant n. Hieruit kunnen we echter niet concluderen dat in dit geval van een frekwentieeffekt sprake is. Het hoogfrekwente woordje "un" ('n) zou dan eveneens beter moeten scoren en dat is in strijd met de resultaten. Bovendien heeft bij alle medeklinkers de klinker aa tot veel hogere correctscores geleid dan de andere klinkers, zodat de hogere score van "na(a)" en "aan" moet worden toegeschreven aan een gunstige invloed van de klinker aa en niet aan een frekwentieeffekt.

Samenvattend concluderen we dat in onze experimenten de herkenning van consonanten in hoofzaak bepaald wordt door fysische eigenschappen van de stimuli en de aantasting daarvan door het systeem van analyse en resynthese. Andere factoren spelen daarbij geen rol van betekenis.

Uit de resultaten blijkt dat het systeem de afzonderlijke medeklinkers in sterk uiteenlopende mate aantast. Daarvoor zijn een aantal

oorzaken aan te wijzen.

De eerste oorzaak ligt bij de modelbeperking dat het brongeluid van de synthese slechts een reeks impulsen of ruis is.

Hierdoor mogen we allereerst verwachten dat stemhebbende wrijfklanken als v en z (gemiddeld over beide 68% correct) worden aangetast. Immers het synthesemodel kent, zoals we in hoofdstuk 2 zagen, geen kombinatie van periodiek en ruisig brongeluid. Frames van spraaksegmenten die zowel periodieke als ruisige componenten bevatten moeten dus geforceerd als stemhebbend òf als stemloos worden geklassificeerd, afhankelijk van twee fysische eigenschappen: energie en eerste autocorrelatie van het spraaksignaal in het analysevenster. De gevonden v→ g(=ch) en v→ f verwisselingen zijn dan te verklaren doordat de stem/stemloosdetector teveel stemloze frames heeft opgeleverd en de v→ w verwisselingen door te weinig stemloze frames. Ook de z→ s verwisselingen kunnen zo worden verklaard door een teveel aan stemloze frames. We zullen in de volgende paragraaf aantonen dat stemhebbende wrijfklanken perceptief sterk verbeterd kunnen worden door de ene helft van de frames stemhebbend en de andere stemloos te maken.

Maar niet alleen stemhebbende wrijfklanken worden aangetast door deze geforceerde indeling in stem/stemloos. Ook delen van periodieke spraaksignalen waar de amplitude laag is vergeleken met de achtergrondruis of met de kwantiseringsruis die bij omzetting van analoog naar digitaal signaal ontstaat, zullen door fouten in de stem/stemloosbeslissing worden aangetast. Zoals we in hoofdstuk 2 hebben vermeld worden signalen met lage amplitude bij voorbaat als stemloos geklassificeerd, tenzij de verhouding tussen eerste autocorrelatie en energie hoog is. Hierdoor zal de resynthese van zwakke stemhebbende spraaksegmenten soms ten onrechte stemloos zijn. Omgekeerd kunnen ook frames ten onrechte stemhebbend zijn, b.v. bij die spraakklanken waarin een ruiscomponent aanwezig is die kort duurt vergeleken met de lengte van het analysevenster. In zo'n geval zal de periodieke komponent overheersen waardoor de resynthese ten onrechte stemhebbend wordt.

Bij de stemhebbende plofklank b kunnen hiermee de gemaakte b→ v (11%) verwisselingen gedeeltelijk worden verklaard. Bij de b is tijdens de opbouwfase van de luchtdruk, vóór het opheffen van de lipafsluiting, de amplitude van het periodieke signaal laag. Na resynthese blijken in bijna alle stimuli wel een of meer frames ten onrechte stemloos te zijn geworden. We hebben eerder opgemerkt dat afwisselend groepjes stemhebbende en stemloze frames perceptief de indruk van een stemhebbende wrijfklank opleveren, waarmee de b→ v verwarringen verklaard kunnen worden.

Na de opbouwfase gaat bij de b het opheffen van de lipafsluiting

vaak gepaard met een kort ruisplofje. Ook dat komt niet altijd ongeschonden door het systeem heen. De stem/stemloosdetector 'ziet' dit ruisplofje vaak niet omdat de periodieke komponent overheerst. Daardoor worden na de opbouwfase alle frames verder stemhebbend en dat kan misschien verklaren waarom er b→w (18%) verwisselingen worden gemaakt.

Ook de approximant h is vaak een kombinatie van ruisig en periodiek geluid. Teveel stemloze frames na resynthese leidt tot h→g(=ch) verwarringen en omgekeerd maken ten onrechte stemhebbende frames verwarring met andere approximanten als l, w en j plausibel.

Een tweede oorzaak van aantasting ligt in de hantering van de 'absolute' spraak/stiltedrempel. In hoofdstuk 2 hebben we vermeld dat beneden een bepaalde energie het signaal binnen het analysevenster als stilte (nul) wordt beschouwd en er geen verdere analyse plaats vindt. Deze drempel maakt dat sommige zachte gedeelten van het spraaksignaal na resynthese nul worden.

Inspectie van de golfvorm heeft laten zien dat deze 'gaten' in de resynthese inderdaad bij enkele consonanten zijn voorgekomen. Dit blijkt te zijn gebeurd in 5 stimuli met de b waarin de golfvorm een lage amplitude heeft tijdens de opbouwfase en in 2 stimuli met de w, waardoor w→r verwisselingen kunnen worden verklaard. Tenslotte is bij de h in de onbeklemtoonde Ce postitie het ruisstootje voor de klinkerinzet te kort geworden, hetgeen verklaren kan waarom er relatief veel h→p verwarringen zijn gemaakt.

Hiermee is er evidentie dat een deel van de foute antwoorden bij de herkenning van plofklanken, stemhebbende wrijfklanken en approximanten toegeschreven kan worden aan het feit dat voor de resynthese het brongeluid geforceerd moet worden ingedeeld in periodiek, ruis of stilte.

Een derde type aantasting kan veroorzaakt worden door afwijkingen in de spectrale omhullende. Doordat het synthesefilter slechts 8 coëfficiënten heeft en de overdrachtsfunktie louter polen bevat kan de omhullende van het inputspectrum slechts met beperkte nauwkeurigheid worden benaderd. Bij nasalen zijn vaak steile hellingen (nulpunten) in het spectrum aanwezig. Om die hellingen nabij zo'n nulpunt vlak te maken zijn meer coëfficiënten van het analysefilter nodig dan bij slappe hellingen, waardoor er minder coëfficiënten beschikbaar zijn om de spectrale toppen elders te compenseren. Kleine verschillen in de ligging daarvan blijven dan na resynthese misschien onvoldoende bewaard om nasalen onderling perceptief goed te kunnen onderscheiden. Of dit beperkte aantal filtercoëfficiënten inderdaad de oorzaak is van veel herkenningsfouten bij nasalen hebben we niet nader onderzocht.

Dit zou kunnen gebeuren door analyse en resynthese met meer coëfficiënten uit te voeren.

Samenvattend blijkt uit de consonantherkenningstest voor normale analyse en resynthese dat het systeem spraaksignalen zodanig aantast dat met name de herkenning van de plofklank b, de stemhebbende wrijfklanken v en z, de nasalen m, n en ng alsmede de approximanten w en h worden bemoelijkt. De belangrijkste oorzaak hiervan ligt bij de modelbeperking die de synthese slechts met periodiek of ruisig brongeluid toelaat en bij fouten in de stem/stemloos- en spraak/stilte beslissingen. Hierbij moet worden opgemerkt dat de analyse volledig automatisch is uitgevoerd. Correctie achteraf van de stem/stemloos- parameter levert voor een aantal consonanten duidelijk verbetering op, zoals we in de volgende paragraaf zullen zien. Dat wijst erop dat een aanzienlijk deel (voor een aantal consonanten meer dan de helft) van de gemaakte verwarringen inderdaad is terug te voeren op de hier genoemde oorzaken. Voor meer kwantitatieve uitspraken hierover is echter meer onderzoek nodig, waarin zowel de stem/stemloos- als de spraak/stiltebeslissing worden gemanipuleerd.

2. Verbetering door stem/stemlooscorrectie

Bij de normale versie NoP is de stem/stemloosbeslissing automatisch tot stand gekomen. Duidelijk hoorbare fouten zijn voor een aantal consonanten via een interaktief programma met 'oor en hand' gecorrigeerd, zodanig dat ze na resynthese perceptief beter overeenkwamen met de oorspronkelijke versie PcM. Bij de p en de h zijn de ten onrechte stemhebbende frames vlak voor de klinkerinzet stemloos gemaakt. Tevens is bij de h een aantal ten onrechte stemloze frames in stemhebbend veranderd. Bij de wrijfklanken v en z zijn, voor zover ze dat niet al waren, de frames van de eerste helft stemloos en de rest stemhebbend gemaakt. Tenslotte zijn bij de d in de (onnatuurlijke) VC positie alle frames na afloop van de klinker van stemhebbend in stemloos veranderd.

De gevolgen van deze correcties zien we in fig. 5.3, links voor de P-versie CoP en rechts voor de A-versie CoA, beide vergeleken met de ongecorrigeerde versie NoP. We zien bij d, v en h een forse afname van het foutenpercentage. Uit de (hier niet weergegeven) verwarringsmatrices blijkt dat bij de v vooral de v->f, v->g en v->w fouten flink zijn afgenomen en bij de h minder h->w, h->j en h->l verwarringen zijn ontstaan.

Ook bij de z leidt correctie tot verbetering, maar minder overtuigend; de standaarddeviatie blijft relatief groot. Nadere uitsplitsing van de scores naar de afzonderlijke posities leert dat stem/stemlooscorrectie voor de z in tussenposities eCV en Vce wèl significante ver-

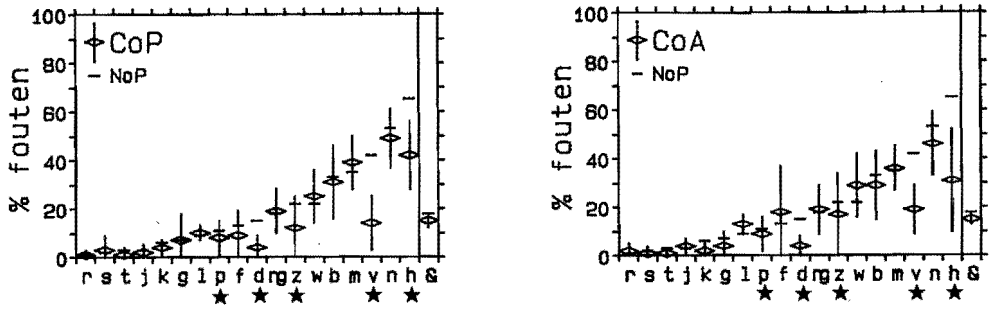

Fig.5.3. Gemiddelde foutenpercentages per consonant voor de stem/stemloosgecorrigeerde versies, links de P-versie CoP, samen met de normale versie NoP en rechts de A-versie CoA, samen met versie NoP. De gecorrigeerde consonanten zijn gemerkt met ★ . Bij & de gemiddelden over alle consonanten.

betering t.o.v. versie NoP heeft opgeleverd, maar dat in de beginposities CV en Ce na correctie juist méér fouten worden gemaakt. Zo wordt de z in de onbeklemtoonde stimulus "ze" in versie NoP foutloos herkend maar na 'correctie' (versie CoP) ontstaan 40% fouten.

Dit illustreert hoe moeilijk het is een goede stem/stemloosindeling uit te voeren. Criteria die in de ene spraakklank tot goede herkenning leiden kunnen bij dezelfde medeklinker in een iets andere situatie juist slechtere herkenning opleveren.

Samenvattend blijkt uit deze resultaten dat stem/stemlooscorrecties tot aanzienlijk betere herkenning van stemhebbende wrijfklanken kunnen leiden, waarmee we hebben aangetoond dat een groot deel van de foute herkenningen bij klanken die zowel periodieke als ruisige componenten bevatten afkomstig is van fouten in de stem/stemloosbeslissing. Deze hangen samen met de modelbeperking die slechts een periodiek of ruisig brongeluid toelaat. Verder onderzoek, met een model waarin het brongeluid ingewikkelder is samengesteld, is hier dan ook op z'n plaats.

3. Invloed van de vaste modelparameters.

In de vorige paragraaf hebben we aangetoond hoe bij een aantal consonanten de herkenning beter wordt door achteraf fouten in de stem/stemloosanalyse te corrigeren. Sommige van die fouten hangen samen met het feit dat voor plofklanken het analysevenster eigenlijk te lang is. Bij de normale analyse is als vensterlengte 25 ms toegepast, hetgeen een compromis is tussen de noodzaak enerzijds het venster kort te houden om bij snel veranderende klanken toch zo goed mogelijk aan de eis van stationair zijn te voldoen en anderzijds om ook bij lage fre-

kwenties van de grondtoon F_0 toch minstens één volle periode te omvatten. Bij plofklanken in vooral prevocale posities duurt het ruisstootje kort vergeleken met het analysevenster. Verkorting van het venster zou dan in principe betere herkenningsscores voor de d moeten opleveren. Daarnaast kunnen we ons afvragen of de relatief slechte herkenning van de klinkerachtige w misschien positief wordt beinvloed door een langer analysevenster.

Behalve de lengte van het analysevenster kan ook de duur van het frame in principe een belangrijke rol spelen bij de verstaanbaarheid van met name plofklanken. De frameduur, dat is de stapgrootte waarmee het venster bij iedere volgende analyseslag wordt opgeschoven, mag niet te groot zijn, teneinde snelle spectrale veranderingen in het spraaksignaal te kunnen volgen. Aan de andere kant moet de frameduur ook niet onnodig klein zijn omdat dat nodeloos veel rekentijd en opslagcapaciteit kost.

Hoe de consonanten herkend worden na analyse en resynthese met andere dan de normale waarden voor analysevenster en frameduur zien we in fig. 5.4 voor een smal (10 ms) resp. breed (60 ms) venster en een korte (4 ms) resp. lange (40 ms) frameduur. De foutenpercentages zijn steeds vergeleken met die van de normale versie NoP (25 ms venster en 8 ms frameduur).

Gemiddeld over alle consonanten blijkt dat zowel kort als lang venster slechtere herkenningsscores opleveren, resp 77% en 72% correct. We zien, in tegenstelling tot de verwachting, bij versie SmA voor de plofklanken geen verbetering t.o.v. versie NoP; bij de b een duidelijke verslechtering en bij de d een dramatische achteruitgang. Alleen bij de v is verbetering te constateren. Voor de versie BrA met breed venster gaan de plofklanken, zoals te verwachten is, flink achteruit maar is evenmin bij de w verbetering te zien.

Bij de korte-frame-versie KoA treedt, zoals verwacht, over de hele linie een lichte vooruitgang op t.o.v. versie NoP, met gemiddeld over alle consonanten 84% correct. De lange frames van versie LaA zijn daarentegen niet alleen nadelig voor de plofklanken (zoals verwacht) maar vooral ook voor de klinkerachtige w en in mindere mate de l.

Discussie

In versie SmA met smal venster is de opvallend sterke toename van fouten bij de stemhebbende plofklanken b en d in eerste instantie onverwacht. Bij de b nemen de in versie NoP gevonden b→w en b→v verwarringen duidelijk af, waaruit we kunnen concluderen dat het plofkarakter nu beter bewaard blijft. Maar er ontstaan met dit korte venster aanzienlijk meer b→p, b→k en b→h verwisselingen. Bij de d treedt een zeer sterke toename op naar maar liefst 39% d→t verwisselingen.

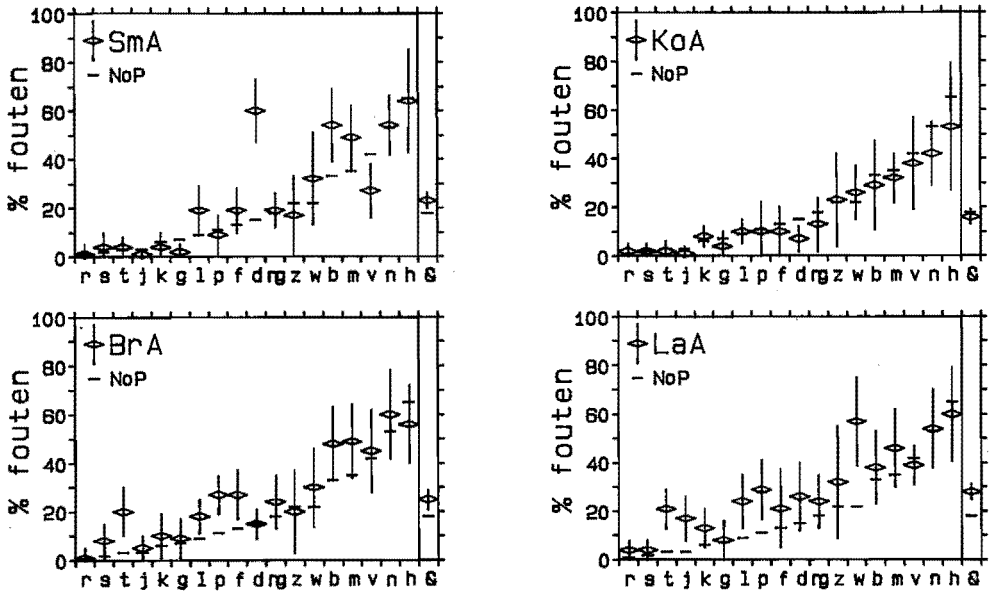

Fig.5.4. Gemiddelde foutenpercentages per consonant voor de versies waarvan de vaste modelparameters zijn gevarieerd. Links boven voor versie SmA met 10 ms (smal) analysevenster, links onder versie BrA met 60 ms (breed) venster. Rechts boven versie KoA met 4 ms (korte) frameduur en rechts onder versie LaA met 40 ms (lange) frameduur. Ook de scores voor de normale versie NoP zijn aangegeven. Bij & de gemiddelden over alle consonanten.

Oorzaak van deze toename moet wel zijn de eerder genoemde foute beslissingen voor stem/stemloos en voor spraak/stilte. Verkorting van het analysevenster leidt ertoe dat in de zachte passages (tijdens de opbouwfase van de druk) van het spraaksignaal veel meer frames ten onrechte stemloos of nul zijn geworden. Dit verklaart waarom de proefpersonen nu meer stemloze plofklanken als p en t antwoorden. Het zonder meer verkorten van het analysevenster is dus geen goede procedure om stemhebbende plofklanken als b en d beter te doen herkennen. Zo'n analyse met kort venster zou op z'n minst gepaard moeten gaan met een aanpassing van zowel stem/stemloos- als spraak/stiltecriterium. Of daarmee dan wel hogere herkenningsscores voor de b en de d bereikt kunnen worden, zonder andere consonanten nadelig te beinvloeden, is hier niet onderzocht en zal uit toekomstig onderzoek moeten blijken.

Dat een smal analysevenster bij de v wel betere scores oplevert dan in de normale versie NoP is eveneens verklaarbaar doordat meer frames stemloos worden geklassificeerd en vaker afwisselingen van groepjes stemhebbende en stemloze frames ontstaan. Daardoor nemen zowel v→w als v→g verwarringen af.

Geheel volgens verwachting levert een breed analysevenster slechtere resultaten op voor de plofklanken. Met name de kort durende p en t hebben te lijden van het uitsmeereffekt van zo'n lang venster. Dat leidt dan tot meer p→b en t→d maar ook tot p→h en p→s verwisselingen. Ook een verdere toename van b→w verwarringen t.o.v. versie NoP kan hieraan worden toegeschreven. Dit kan overigens een aanwijzing zijn dat ook bij de normale versie NoP, waar veel b→w verwarringen worden gemaakt, het analysevenster eigenlijk te lang is voor stemhebbende plofklanken.

Het brede analysevenster levert ook geen verbetering voor de klinkerachtige w. De responsies blijven sterk verspreid met een lichte concentratie naar w→v en w→r verwarringen, net als bij versie NoP. Dit is misschien gedeeltelijk toe te schrijven aan stem/stemloosfouten bij lage amplitude in de VC en CV posities.

De korte frameduur van 4 ms (versie KoA) geeft, zoals verwacht, over de hele linie een lichte verbetering. Eveneens in de lijn der verwachting ligt de toename van foute herkenningen bij de lange frameduur van 40 ms in versie LaA. Snelle veranderingen worden nu niet meer snel genoeg gevolgd en dat levert vooral bij de stemloze plofklanken p en t slechtere herkenningsscores op. De golfvorm is na resynthese aangetast, doordat vanwege de grote frameduur vooral prevocale ruisplofjes vaak worden overgeslagen, hetgeen leidt tot p→b en t→d verwisselingen. En wanneer het ruisplofje wel in het analysevenster aanwezig is en door de stem/stemloosdetector 'gezien' wordt, maakt de grote frameperiode de duur ervan veel te lang, hetgeen leidt tot p→k, t→k, t→h en k→g(=ch) verwisselingen. De stemhebbende plofklanken b en d zijn in de NoP versie al aangetast; verlenging van de frames maakt de achteruitgang naar verhouding minder sterk.

Opvallend sterk is de teruggang in herkenning bij de klinkerachtigen j, l en vooral de w. De grote frameduur leidt tot j→l, l→r, l→w, w→r en w→l verwisselingen, waarvoor waarschijnlijk verlies van details in het spectrale verloop verantwoordelijk is. Hierover hebben we geen verdere evidentie.

Samenvattend volgt uit de resultaten met gevarieerde modelparameters:

- Een korte frameduur van 4 ms leidt over de hele linie tot een lichte verbetering in consonantherkenning tov de normale frameduur van 8 ms.
- Verlenging van de frameduur tot 40 ms levert niet alleen slechtere herkenning voor de plofklanken, maar ook voor de klinkerachtige approximanten l, j en w.
- Stemhebbende plofklanken d en vooral b, die na analyse met een normaal venster van 25 ms relatief slecht herkend worden, scoren in tegenstelling tot de verwachting, ook slechter bij een kort venster

van 10 ms. Reden hiervan is dat het korte venster leidt tot meer stem/stemloosfouten en tot meer spraak/stiltefouten.
- De klinkerachtige approximant w, die bij een normaal venster van 25 ms relatief slecht wordt herkend, scoort niet beter door het analysevenster te verlengen tot 60 ms.

4. Resonerend maken van alle deelfilters.

In hoofdstuk 3 hebben we gezien dat verandering van het analysefilter in louter toegevoegd complexe poolparen (resonanties) het oorspronkelijke energiespectrum nauwelijks wijzigt. In de gevallen waarbij wel veranderingen ontstaan kunnen die in principe van invloed zijn op de verstaanbaarheid van afzonderlijke medeklinkers. Dat zou dan moeten blijken door de resultaten van de versies NoA en CoA te vergelijken met die van de versies NoP en CoP, waarvan bij de analyse alle deelfilters resonerend zijn gemaakt.

In fig. 5.5 zijn de foutenpercentages voor de normale en de stem/stemloosgecorrigeerde A-versies NoA en CoA uitgezet en vergeleken met resp. de normale P en de gecorrigeerde P-versie NoP en CoP. We zien dat de verschillen tussen de A- en de P-versies voor vrijwel alle medeklinkers klein zijn en er is geen aanwijzing dat de P-versies significant slechter scoren dan de A-versies.

Een eventuele nadelige invloed van het resonerend maken van alle deelfilters van het analysefilter zou vooral bij nasalen tot uiting moeten komen. Immers in het energiespectrum van nasalen komen vaak 'nulpunten' of steile hellingen voor, waardoor deelfilters vaak reële poolparen in de overdrachtsfunktie bevatten en niet resonerend zijn.

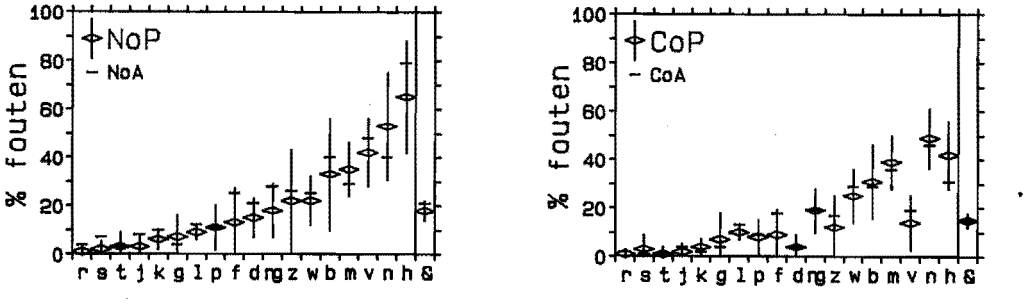

Fig.5.5. Gemiddelde foutenpercentages per consonant voor de P-versies waarvan alle deelfilters resonerend zijn gemaakt, vergeleken met de overeenkomstige oorspronkelijke A-versies. Links voor versie NoP en NoA, rechts voor CoP en CoA. Bij & de gemiddelden over alle consonanten.

Deze niet-resonerende deelfilters zijn in de versies NoP en CoP 'geforceerd' resonerend gemaakt, waardoor de verstaanbaarheid van de nasalen dus minder zou kunnen worden.

Uit de hier gevonden resultaten blijken daarvoor geen duidelijke aanwijzingen. Wel worden bij de m en de n voor de P-versies met louter resonanties gemiddeld wat meer fouten gemaakt dan in de A-versies (met vooral ng als antwoord) maar de toename valt ruim binnen de standaarddeviaties. Het resonerend maken van de deelfilters heeft dus perceptief geen significant nadelige gevolgen.

5.2.3. Literatuur

In deze paragraaf zullen we enkele resultaten van onze verstaanbaarheidsproeven vergelijken met gegevens uit de literatuur.

Voor PcM gecodeerde spraak zijn een drietal studies relevant.

Goodman et al (1978) hebben met een consonant herkenningstest (CRT) 18 'enkelletterige' beginconsonanten uit het brits engels getest met CVC combinaties die, evenals in onze test, deels bestaande, deels onzinwoordjes vormen. Deze werden aangeboden met in de luisterkamer een achtergrondruis van 50 dB(A). De "unprocessed source tape" (niet nader vermeld wat dat precies is) leverde een gemiddelde correctscore over alle consonanten van 94%. Relatief laag scoorden de p (82%) en de v (79%). Voor PcM spraak, 3 kHz bandgefilterd en met 6.4 kHz en 8 bits gedigitaliseerd, zakt de gemiddelde score naar 78% correct, met lage waarden voor de p (37%), de v (48%), de s (63%) en de z (51%). Beide resultaten, zowel voor de "unprocessed" spraak als de PcM spraak liggen beduidend lager dan voor onze spraak die met 8 kHz en 12 bits is gecodeerd, maar de uitkomsten zijn moeilijk in absolute zin met elkaar te vergelijken. Het is vrij aannemelijk dat de achtergrondruis de score bij Goodman et al (1978) nadelig heeft beinvloed.

In een studie van Nye en Gaitenby (1973) is spraak getest die met onze PcM versie beter vergelijkbaar is (ook 8 kHz PcM, 3.5 kHz gefilterd), echter met een gemodificeerde rijmtest (MRT). Proefpersonen kregen daarbij bestaande engelse woordjes aangeboden, waarvan begin- of eindconsonant gekozen moest worden uit 6 antwoordalternatieven. Dat leverde voor de beginconsonanten 98% en voor de eindconsonanten gemiddeld 97% correcte score op. Hoewel de methode nogal verschilt van de onze komen zowel spraaksoort als score aardig overeen.

Pols (1979) heeft in een evaluatie van het 'diade'-synthesesysteem van Olive (1976-1979) o.a. 4 kHz bandgefilterde, met 10 kHz en 12 bits gedigitaliseerde spraak gebruikt. Met nonsens CVC woordjes werden 22 consonanten uit het amerikaans engels getest, waaronder ook 'dubbel-

letterige' zoals bv. de beginconsonant uit "thief" en uit "that". Dit leverde een correctscore op van 93%, gemiddeld over begin- en eindconsonant. Daarbij scoorden de dubbelletterige duidelijk lager dan de overige. Als we in Pols' resultaten alleen de consonanten meerekenen die ook in onze test zijn gebruikt, en als we uit onze resultaten de door Pols niet geteste g=ch weglaten komen de scores voor Pols op 96% en voor ons op 97% correct, hetgeen dan aardig met elkaar overeenstemt.

Voor geresynthetiseerde spraak zijn vrijwel geen literatuurgegevens beschikbaar die vergelijkbaar zijn met onze consonant-herkenningstest. Het merendeel van dit soort spraak is geëvalueerd met de diagnostische rijmtest (DRT), voorgesteld door Fairbanks (1958) en verder ontwikkeld door Voiers (1977). In deze test krijgen proefpersonen 96 CVC woordparen visueel aangeboden, waarvan alleen de beide beginconsonanten auditief van elkaar verschillen met betrekking tot een van de 6 foneemkenmerken: 'sustention' (aanhoudend): then - den; 'sibilation' (sissend): joe - go; 'graveness' (donker): fin - thin; 'compactness' (licht): gill - dill; 'nasality' (nasaal): moss - boss en 'voicing' (stemhebbend): bean - pean. Taak van de proefpersoon is om voor ieder geresynthetiseerd woordje aan te strepen welke van de twee mogelijke is gehoord.

De uitkomsten van deze test, gecorrigeerd voor raden, gemiddeld over de 6 kenmerken en de luisteraars, zijn dan afhankelijk van spreker, details van het synthesesysteem en het aantal gebruikte filtercoefficienten. De hoogste score vinden we bij Keeler et al (1976). Spraak die is gecodeerd met 12 filterparameters leverde bij een spreker 92% correctscores op. Smith et al (1981) komen iets lager uit, zij vinden voor 12 filterparameterspraak van een tenor en een bariton resp 89% en 88% correct. Voor de afzonderlijke attributen kunnen de scores aanzienlijk onderling uiteenlopen. 'Sustention' en 'graveness' scoren soms maar 60 of 70% correct, terwijl 'nasality', 'sibilation' en 'voicing' vaak meer dan 90% correct scoren (Keeler et al, 1976; Smith et al, 1981). We gaan hier echter de deelscores van de afzonderlijke DRT-kenmerken niet verder vergelijken met onze consonantscores omdat zowel het principe van beide methodes alsook het gebruikte stimulusmateriaal daarvoor te ver uiteenlopen, terwijl de door ons uiteraard niet toegepaste engelse consonanten niet uit de literatuurresultaten van zo'n DRT geisoleerd kunnen worden.

In genoemde studie van Keeler et al (1976) worden ook resultaten van een gemodificeerde rijmtest MRT vermeld, waarbij gemiddeld over begin- en eindconsonanten voor geresynthetiseerde spraak (waarvan het filter uit 12 parameters bestaat) een correctscore van 88% is gevonden. Maar zoals we al bij de bespreking van de resultaten voor PcM spraak hebben

opgemerkt, is ook deze MRT niet goed vergelijkbaar met de door ons toegepaste consonantverstaanbaarheidstest.

We keren daarom weer terug naar genoemde studie van Pols (1979). De door hem geteste geresynthetiseerde spraak (12e orde filter, stem/stemloos gecorrigeerd en mèt toonhoogte) leverde gemiddeld over begin- en eindconsonant 86% correctscore op. Wanneer we dh zh en th uitsluiten wordt dit 87%. Onze resultaten zonder g=ch komen gemiddeld uit op 85% correct, en dat is ook weer aardig met elkaar in overeenstemming. Dat neemt niet weg dat voor de afzonderlijke consonanten, gemiddeld over begin- en eindpositie wel verschillen zijn te constateren. In ons systeem scoort de d 20% en de f 11% hoger dan bij Pols. De nasalen als geheel genomen doen het bij ons echter 10% slechter, waarbij in detail onze m en n resp 14% en 36% lager scoren en de ng 13% hoger uit komt dan in Pols' studie. Het is denkbaar dat het feit dat onze analyse en resynthese met 4 parameters minder zijn uitgevoerd toch enige invloed heeft op de nasalen. Aanwijzingen daarvoor hebben we ook al gezien bij de fysische evaluatie in par. 4.2.2. De w scoort in ons systeem 20% slechter, waarbij wel moet worden opgemerkt dat Pols de w in eindpositie niet heeft getest, terwijl die in onze resultaten duidelijk lager scoort dan in beginpositie en daarmee de gemiddelde score dus verlaagt. Ook wat betreft begin- versus eindpositie van andere consonanten zijn er overeenkomsten tussen Pols' en onze resultaten. Zo is in beide posities de b erg zwak en levert in de beginpositie de b nog meer fouten op dan in eindpositie. Ook bij Pols is de m in eindpositie duidelijk veel slechter geidentificeerd dan in de beginpositie, terwijl dat verschil tussen beide systemen voor de n veel kleiner is. In beide systemen tenslotte is de h duidelijk zwak.

Gelet op het feit dat in onze test ook de onbeklemtoonde Ce, eCV en VCe posities zijn getest en dat onze spraak met 8 in plaats van 12 filtercoefficienten is beschreven, is er toch een redelijke overeenkomst te constateren tussen beide studies, zowel voor de PcM spraak als de met beide systemen geresynthetiseerde spraak.

5.2.4. Samenvatting

Uit de consonantherkeningstest, uitgevoerd voor 9 versies geresynthetiseerde spraak van één spreker, kunnen we de belangrijkste resultaten als volgt samenvatten.

- Automatische analyse en resynthese met normale modelparameters leveren 16% vermindering in consonantherkenning op; van 98% naar 82% correct, gemiddeld over alle geteste consonanten en 8 luisteraars.

Aangetast worden vooral de stemhebbende plofklank b, de stemhebbende wrijfklanken v en z, de nasalen m, n en ng en de approximanten w en h.

- Deze aantastingen kunnen grotendeels worden toegeschreven aan een drietal beperkingen in het analyse-resynthesesysteem. Ten eerste worden bij de resynthese als bron slechts òf periodieke impulsen òf ruis òf stilte toegepast. Daardoor wordt de herkenning van klanken met zowel periodieke als ruisige componenten, zoals b, v, z en h bemoeilijkt. Ten tweede wordt de analyse met een vaste vensterlengte van 25 ms uitgevoerd. Dat kan leiden tot stem/stemloos en spraak/stiltefouten en draagt bij tot slechtere herkenning van b en h. Ten derde bestaat het synthesefilter uit louter polen. De spectrale omhullende van nasalen (nulpunten) blijft daardoor vermoedelijk niet nauwkeurig genoeg bewaard om een goede herkenning te verzekeren.
- Verkorting van het analysevenster t.o.v. de normale waarde, om daarmee betere herkenning van de stemhebbende plofklanken b en d te bereiken, heeft geen zin zolang niet tevens stem/stemloos en spraak-/stiltebepaling worden aangepast.
- De procedure waarmee het synthesefilter in louter resonerende deelfilters wordt omgerekend heeft geen significant nadelige invloed op de herkenning van medeklinkers, ook niet op nasalen, die hiervoor in principe het gevoeligst zijn.
- Nader onderzoek zal moeten uitwijzen waarom de nasalen en de klinkerachtige w relatief slecht worden herkend. Daarbij valt in eerste instantie te denken aan uitbreiding van het aantal filtercoëfficiënten om te bezien of het verlies van details in het spectrale verloop hiervoor verantwoordelijk is.

5.3. CONCLUSIES EN NABESCHOUWING

We willen hier voorop stellen dat de gevonden resultaten van de consonantherkenningstest betrekking hebben op slechts één spreker, dezelfde waarmee ook de spraakinterferentietest is uitgevoerd. Deze stem komt zeker niet als beste door het analyse-resynthesesysteem heen. Hij is destijds gekozen vanwege zijn geoefende uitspraak. Andere stemmen leveren, blijkens informele luisterproeven met lopende spraak, vaak aanzienlijk minder verschil op tussen resynthese en oorspronkelijke spraak. Het is dan ook heel goed mogelijk dat het systeem de consonantverstaanbaarheid van andere spreekstemmen minder aantast. Daarover hebben we nu nog geen formele gegevens; verdere evaluatie van het systeem, ook voor uiteenlopende vrouwen- en kinderstemmen, zal moeten uitwijzen of en in hoeverre de resultaten ook daarvoor gelden.

Overigens betekent de gevonden afname in consonantherkenbaarheid van gemiddeld 98% naar 82% correct (voor automatische analyse en resynthese), dat de articulatiescores voor woorden en zinnen door het systeem slechts weinig worden aangetast. Zoals ook uit informele luisterproeven met lopende spraak is gebleken blijft de zinsherkenning vrijwel 100%. Lopende spraak bestaat uiteraard niet alleen uit consonanten en de luisteraar kan door de samenhang van woorden en woorddelen de herkenning tot een goed einde brengen. Ook uit de eerder uitgevoerde spraakinterferentietest (Vogten, 1980) kunnen we al opmaken dat de herkenning van afzonderlijke woorden in zinsstruktuur nauwelijks door het systeem wordt aangetast.

Dit betekent niet dat we daarmee volledig aan de in hoofdstuk 0 gestelde eisen hebben beantwoord. Voor een aantal spreekstemmen treden perceptief soms nog flinke verschillen op tussen origineel en resynthese. Voor verdere verbetering van het analyse-resynthesesysteem zal het onderzoek vooral gericht moeten zijn op toepassing van een meer realistisch bronsignaal bij de synthese. We hebben in hoofdstuk 3 gezien dat het restsignaal van de analyse het 'ideale' brongeluid vormt; samen met het synthesefilter kunnen we hieruit het oorspronkelijke spraakgeluid weer exact reconstrueren. In het systeem wordt dit restsignaal, dat bijna evenveel informatie bevat als het oorspronkelijke spraaksignaal, vervangen door periodieke impulsen, witte ruis of stilte. Daarmee gaat dus veel informatie verloren. Voor lopende spraak is dat, afhankelijk van de spreekstem, perceptief vaak toelaatbaar maar wanneer zwaardere eisen aan het systeem worden gesteld kan deze reduktie van informatie te drastisch zijn. Vooral de spraakklanken die niet eenduidig zijn in te delen in periodiek of ruis, omdat ze beide componenten bevatten, worden nu nog aangetast door het systeem. Toepassing van een ingewikkelder samengesteld brongeluid, waarin meer informatie over het restsignaal bewaard blijft, zal de perceptieve overeenkomst tussen input en resynthese vooral voor deze klanken verder vergroten (Atal en David 1979, Atal en Remde 1982).

Voor de evaluatie van synthetische spraak blijkt de consonantherkenningstest een gevoelige en geschikte methode om specifieke tekortkomingen van dit type spraak vast te stellen. Snellere uitvoering kan in principe bereikt worden door de test te beperken tot 'gevoelige' medeklinkers als stemhebbende plofklanken, stemhebbende wrijfklanken, nasalen en enkele approximanten als w en h, in kombinatie met vooral de klinkers ie en oe.

Een bezwaar van deze (en veel andere) verstaanbaarheidstests is dat daaruit in het geheel niet blijkt of luisteraars bij de resynthese

meer moeite doen om het gehoorde te verstaan dan bij de originele spraak. Het lijkt heel aannemelijk dat een aantasting van bepaalde medeklinkers door het systeem tot gevolg heeft dat de luisteraar meer beroep moet doen op redundantie in de spraak en dat woorden daardoor later herkend worden dan bij de originele spraak (Nooteboom, 1982). Aanwijzingen hiervoor zijn gevonden in de recent uitgevoerde 'groeiwoord'-herkenningstest, waarin gebruik wordt gemaakt van het feit dat meerlettergrepige woorden vaak al herkend worden nog vóór het woord kompleet ten gehore is gebracht (Marslen-Wilson, 1978). Proefpersonen krijgen een kort onherkenbaar beginfragment aangeboden waaraan stap voor stap een segment (klinker of medeklinker) wordt toegevoegd en moeten dan raden van welk woord dat fragment afkomstig is. Nooteboom en Doodeman (1982, 1983) hebben deze aangroeimetingen van Grosjean (1980) toegepast op een 40-tal woorden, uitgesproken door dezelfde spreker als bij onze test voor consonantherkenning. Zij vinden dat voor de resynthese de herkenning ruwweg 1 segment later tot stand komt dan bij de oorspronkelijke PcM-spraak. Naarmate de spraak slechter is blijken er ook meer segmenten nodig te zijn alvorens de herkenning tot stand komt.

Een voordeel van deze groeiwoord-herkenningstest is dat het testmateriaal dicht bij gewone spraak aansluit maar dat geen storende spraak of ruis hoeft te worden toegevoegd om een goede discriminatie tussen verschillende spraaksoorten te verkrijgen. Deze test lijkt dan ook eveneens geschikt voor verdere evaluatie van het analyse-resynthesesysteem.

6 TOEPASSINGEN VAN HET SYSTEEM

Voor spraakonderzoek is onder meer vereist dat het spraaksignaal manipuleerbaar is. Dat wil zeggen dat afzonderlijke fysische eigenschappen beheersbaar zijn ('t Hart ea 1981/1982), om zo hun belang voor de spraakperceptie te kunnen bestuderen. Hier ligt dan ook een belangrijke toepassing van het ontwikkelde systeem voor analyse en resynthese. Immers hiermee wordt spraak geanalyseerd in termen van fysische parameters die nauw samenhangen met belangrijke kenmerken in de spraakperceptie. Die afzonderlijke parameters zijn onafhankelijk te variëren en na resynthese kan het effekt van zo'n verandering op de spraakperceptie worden beluisterd. Hiernaast is onderzoek aan spraak voor een groot deel ook praktijkgericht en wel op de ontwikkeling van apparatuur voor spraakuitgifte. Daarbij wordt met name nagegaan hoe de ingewikkelde struktuur van het spraaksignaal kan worden vereenvoudigd door perceptief onbelangrijke details in het verloop van die parameters weg te laten. De daarmee gepaard gaande bezuiniging is voor de praktijk belangrijk omdat zelfs bij eenvoudige systemen voor uitgifte van bv. direkt toegankelijke gesproken meldingen en waarschuwingen al gauw enkele tientallen seconden spraak moeten kunnen worden opgeslagen. Bij 'normale' PcM golfvormcodering zou dat een geheugenomvang van enkele miljoenen bits vereisen. Wanneer we echter de spraak eerst analyseren en daarna de parameters zuinig coderen kan ruwweg een faktor 100 worden bezuinigd op de vereiste geheugenomvang. Daarmee kunnen dan opslag en resynthese worden ondergebracht in enkele chips, zodat kleine en zeer robuuste uitgiftesystemen zijn te realiseren.

Een beperking van dit soort systemen is evenwel dat in principe iedere uiting vooraf ingesproken en geanalyseerd moet worden, ongeveer op de wijze zoals ze later moet worden uitgesproken door de apparatuur. Weliswaar kan door het aanbrengen van ruime pauzes tussen variabele en vaste stukken van de boodschap enige variatie worden aangebracht maar veel verder kunnen we hiemee niet gaan. Veel flexibeler en krachtiger zouden systemen zijn waarmee we iedere willekeurige boodschap kunnen samenstellen door losse klanken aaneen te rijgen, net zoals teksten worden gevormd door aaneenschakeling van letters, cijfers en leestekens. Probleem bij het zonder meer aaneenrijgen van b.v. fonemen is echter dat dit zeer onnatuurlijk klinkende spraak oplevert, die meestal zelfs onverstaanbaar is. Om tot verstaanbare en vloeiende spraak te komen moeten de parameters zoals grondtoon, amplitude en duur van de spraaksegmenten worden aangepast aan de omgeving waarin ze in de zin worden uitgesproken. Om dat automatisch te kunnen doen zijn regels nodig. Zowel voor het vinden van die regels, alsook voor het vinden van de beste bouwstenen, het genereren, selecteren en

volgens die regels aaneenrijgen wordt ons analyse-resynthesesysteem intensief gebruikt.

We zullen in dit hoofdstuk van de hierboven genoemde toepassingen een aantal voorbeelden bespreken. In de eerste paragraaf zullen we laten zien welke mogelijkheden het systeem biedt bij het spraakonderzoek zelf en gaan we wat dieper in op één aspect daarvan: het intonatieonderzoek. Dan geven we in par. 6.2 een voorbeeld van een praktische toepassing van het systeem voor zuinige codering en opslag van spraak. In de laatste paragraaf komt dan de toepassing van het systeem aan de orde bij de ontwikkeling van een spraakuitgiftesysteem dat is gebaseerd op difoonconcatenatie en waarmee nieuwe spraakuitingen kunnen worden geproduceerd, die nooit eerder als zodanig door een menselijke stem zijn uitgesproken.

6.1 TOEPASSING IN HET SPRAAKONDERZOEK

In voorgaande hoofdstukken hebben we gezien hoe de golfvorm van het spraaksignaal, afkomstig van microfoon of bandrecorder, in digitale vorm wordt opgeslagen in het schijvengeheugen van de computer en bewaard in files die we hier verder N-files zullen noemen. Uit zo'n N-file worden in de (standaard) analyse in stapjes van 10 ms steeds de 3 bronparameters en de 10 filterparameters (a- of pq- coëfficiënten) berekend, die op hun beurt worden opgeslagen in een file die we hier verder aanduiden met A/P-file. Bij de resynthese daarvan ontstaat vervolgens een nieuwe N-file die net als de oorspronkelijke N-file na

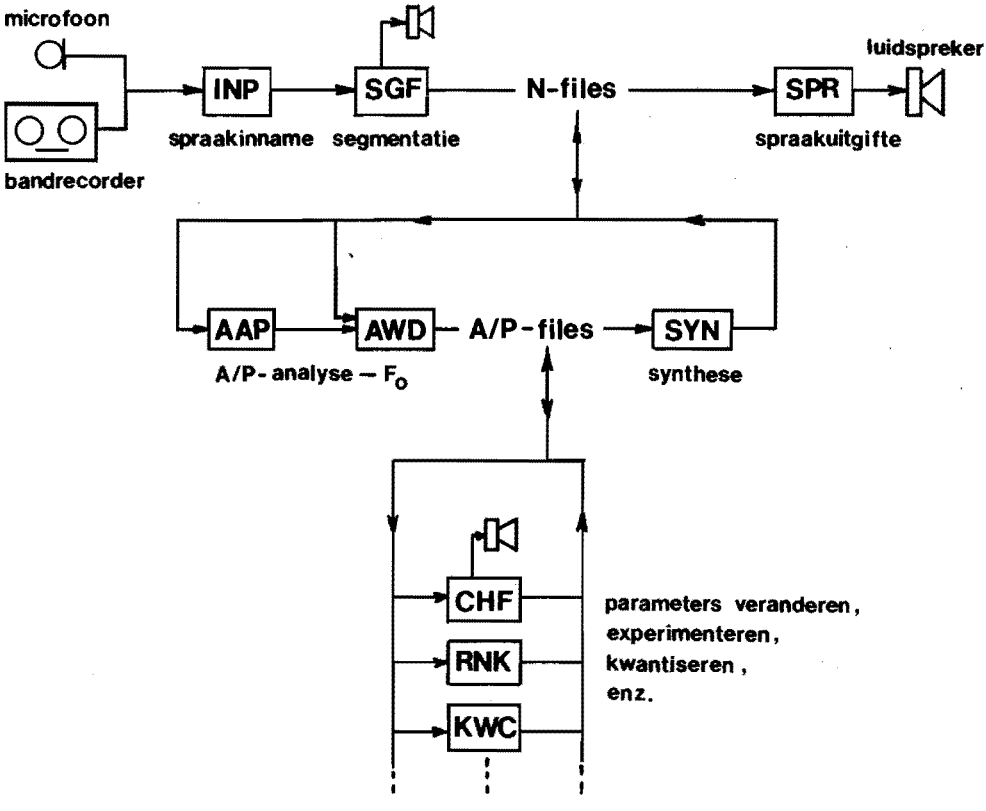

Fig.6.1. Schematisch overzicht van enkele essentiele programma's van het analyse-resynthesesysteem voor (lange) files waarin spraakuitingen zijn opgeslagen. In N-files is de gedigitaliseerde golfvorm gecodeerd. De zgn A/P-files bestaan uit een aaneenschakeling van frames waarin de (meestal 13) modelparameters zijn opgeslagen. Met de (interaktieve) programma's SGF en CHF kunnen ook fragmenten direkt ten gehore worden gebracht, hier aangegeven door de luidspreker.

digitaal-analoogomzetting beluisterd kan worden via koptelefoon of luidspreker. Deze opeenvolgende stappen zijn schematisch in fig. 6.1 weergegeven.

A/P-files nemen bij het spraakonderzoek in ons systeem een centrale plaats in. Het zijn deze files waarin alle parameterwaarden van de analyse naar wens kunnen worden veranderd en eventueel opgeslagen in nieuwe A/P-files, waarna resynthese het effekt van zo'n wijziging op de spraakperceptie laat horen.

Voor dagelijks gebruik in het spraakonderzoek is een 50-tal programma's ontwikkeld die allerlei bewerkingen, variërend van spraakinname in de computer tot en met de uitgifte, grotendeels automatisch verzorgen. Ze worden door de gebruiker van het systeem via de terminal aangeroepen door intikken van een kort commando en een of meer namen van de files die voor die betreffende programma als input of als output dienen. Zo zien we in fig. 6.1 de aanduiding AAP voor het commando waarmee een reeks N-files volledig automatisch wordt geanalyseerd in a- of pq-parameters, met als resultaat een reeks A/P-files. Met het programma SYN worden deze gesynthetiseerd in nieuwe N-files die dan samen met de oorspronkelijke N-files, met het programma SPR, beluisterd kunnen worden. Daarnaast zijn er ook programma's ontwikkeld waarmee op interaktieve wijze grafisch willekeurige segmenten van N- of A/P-files kunnen worden aangewezen, bekeken, beluisterd, bewerkt en weggeschreven in dochter-files. Voor een complete lijst van programma's, een beschrijving van de vele mogelijkheden en een gebruiksaanwijzing zij verwezen naar Vogten (1983).

6.1.1. Interaktief parameters wijzigen

Bij de interaktief werkende programma's wordt gebruik gemaakt van een terminal waarmee de inhoud van b.v. A/P-files (het resogram) grafisch wordt weergegeven (grafische output), en waarmee tevens met behulp van een wijzer (cursor) willekeurige segmenten van files zijn aan te wijzen (grafische input) en parameterwaarden snel ingevoerd kunnen worden in de computer.

Een voorbeeld van zo'n interaktief werkend programma is CHF, waarmee een groot aantal bewerkingen op de parameters van een A/P-file kunnen worden uitgevoerd. Na het intikken van het commando CHF en de filenaam krijgt de gebruiker het komplete resogram in beeld van een (vrij te kiezen) fragment van 100 frames. Bij een standaard frameduur van 10 ms is dat dus 1 seconde spraak. Door met b.v. duimwielen de x-positie van de cursor in te stellen kan ieder gewenst frame worden aangewezen, waarna er vele mogelijkheden zijn.

Te beginnen met het door de cursor aangewezen frame kan, door indrukken van een van de terminaltoetsen een korter of langer stuk van de file worden geresynthetiseerd en ten gehore gebracht. Dat kan dan direct (auditief) worden vergeleken met hetzelfde fragment van de oorspronkelijke N-file of, afhankelijk van de ingedrukte toets, dat van een derde file.

Verder kan met de cursor het begin- en eindframe van een willekeurig segment worden gespecificeerd. Binnen dit segment kunnen vervolgens parameters b.v lineair worden geinterpoleerd tussen hetzij de bestaande waarden van begin- en eindframe, hetzij nieuw te specificeren parameterwaarden. Deze nieuwe waarden worden gegeven door de ingestelde y-positie van de cursor of door intikken in getalvorm voor het geval een grotere precisie is gewenst. Ook kunnen een of meer parameters binnen het aangegeven segment met een nader in te voeren konstante worden vermenigvuldigd. In de praktijk worden deze mogelijkheden vooral toegepast om eventuele fouten in de analyseresultaten van de grondtoon F_0 en/of de stemloosparameter VUV snel te corrigeren.

Behalve het veranderen van de parameterwaarden in het gespecificeerde segment (dat ook uit de hele file kan bestaan) en het wegschrijven daarvan in een nieuwe file, kan ook de tijdsduur van zo'n segment naar willekeur worden veranderd. We hebben in hfdst 2 gezien dat de standaardanalyse wordt uitgevoerd in vaste stappen van 10 ms. Niets weerhoudt ons er echter van om bij de resynthese van bepaalde segmenten een andere frameduur te kiezen. Dat kan heel eenvoudig door die frames van een extra parameter te voorzien, die de 'lokale' frameduur aangeeft. Door deze bij de resynthese dan te laten prevaleren boven de frameduur die bij de analyse is toegepast, kan ieder aangewezen spraaksegment willekeurig versneld of vertraagd worden weergegeven. Daarbij blijven uiteraard toonhoogte en spectrale samenstelling onveranderd.

Het effekt van deze veranderingen op de geresynthetiserde spraak kan door indrukken van de betreffende toets op de terminal direct worden beluisterd. Dat leent zich uiteraard zeer goed voor talloze bewerkingen.

Zo kan b.v. 1) volstrekt monotone spraak worden gemaakt door F_0 vast te zetten, 2) een gestileerde toonhoogtecontour worden aangebracht door (op log schaal) lineaire interpolatie tussen een aantal F_0-waarden, 3) een falsetstem bij mannenspraak worden gesimuleerd door de gehele gemeten F_0-contour, via vermenigvuldiging met b.v. een factor 2, omhoog te schuiven, 4) 'hese fluisterspraak' worden nagebootst door bij alle frames de stem/stemloosparameter op stemloos te fixeren. Ook kan worden gedemonstreerd dat hogere formanten minder belangrijk zijn voor het verstaan van de spraak dan lagere, door b.v. voor F4 en F5 een konstante waarde te kiezen en de resynthese te

beluisteren in vergelijking met die van dezelfde file waarin F_1 en F_2 constant zijn gemaakt. Verder kan direct worden gedemonstreerd hoe b.v. van een woordje de klinker "ie" kan overgaan in "oe" door alleen de F_2 parameter in het frekwentiegebied omlaag te verschuiven, of "uu" kan worden veranderd in "ie" door F_2 naar hogere frekenties te verschuiven. Door vermenigvuldiging van F_0 met een factor 2 en de FB-waarden met 1.15 kunnen we een mannenstem laten overgaan in een vrouwenstem. Of omgekeerd vanuit een vrouwenstem een mannenstem simuleren door F_0 met 0.5 en de FB-parameters met 0.85 te vermenigvuldigen.

Het zou te ver voeren om hier alle mogelijkheden van dit universele programma CHF op te sommen. Voor een gedetailleerde handleiding verwijzen we naar Vogten (1983). Dit programma is hoofdzakelijk ontwikkeld voor experimenteer- en demonstratiedoeleinden. Hiermee zijn alle modelparameters snel en handig te veranderen en zijn die veranderingen direct grafisch zichtbaar te maken, terwijl de gevolgen voor de spraakperceptie onmiddellijk hoorbaar worden gemaakt.

In de volgende paragraaf zullen we een voorbeeld geven waarbij het analyse-resynthesesysteem wordt gebruikt bij het onderzoek aan één van deze parameters: de toonhoogte.

6.1.2. Intonatie-onderzoek

In het intonatieonderzoek wordt bestudeerd welke invloed het verloop van de grondtoonfrekwentie F_0 heeft op het waargenomen toonhoogteverloop (de zinsmelodie) van een spraakuiting. Daarbij worden onder meer de regels opgespoord waaraan de F_0-contour moet voldoen om goed verstaanbare en natuurlijk klinkende spraak te verkrijgen. Ook hiervoor is het analyse-resynthesesysteem een nuttig stuk gereedschap ('t Hart et al 1982). Het analyseresultaat, met een meestal vrij grillig verloop van F_0 in de tijd, kan hiermee geheel naar wens van de onderzoeker worden vereenvoudigd. Via het beluisteren van de dan geresynthetiseerde spraak kan worden vastgesteld hoe ver die vereenvoudiging mag gaan zonder de perceptieve overeenkomst met de oorspronkelijke resynthese te verliezen. Dan blijkt dat sommige F_0-bewegingen wel perceptief relevant zijn en b.v. een toonhoogteaccent verlenen aan beklemtoonde (delen van) woorden. Andere F_0-bewegingen zijn daarentegen niet van belang voor de waargenomen intonatie en kunnen dan ook worden gladgestreken zonder de perceptieve indruk te verstoren. We spreken dan van een 'close-copy' stilering.

Een voorbeeld hiervan is in fig. 6.2 weergegeven. Een resynthese met de oorspronkelijk gemeten F_0-contour van de zin "vannacht is de vorst

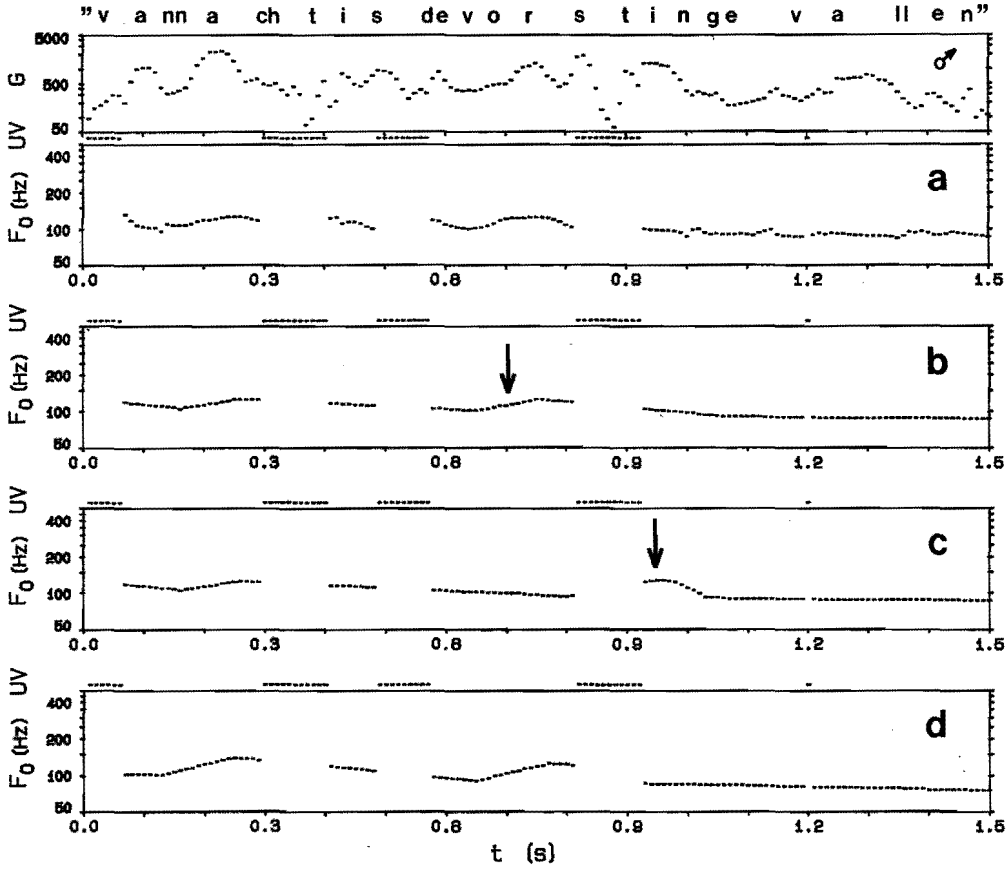

Fig.6.2. Boven: bronparameters van de zin "vannacht is de vorst ingevallen" met in (a) de oorspronkelijk gemeten toonhoogtecontour. Daaronder in (b) een close-copy stilering die na resynthese perceptief niet is te onderscheiden van (a). In (c) is de oorspronkelijke F_0-beweging bij "vorst" (pijl) opgeschoven naar "in", waardoor de zin als "vannacht is de vorstin gevallen" wordt waargenomen. In (d) een stilering volgens de intonatie-grammatica van 't Hart et al (1973, 1975).

---

ingevallen" van fig. 6.2a is voor gewone (niet speciaal getrainde) luisteraars niet te onderscheiden van die met de sterk vereenvoudigde contour van fig. 6.2b. We zien hier geïllustreerd hoe perceptief belangrijke, dwz accentverlenende bewegingen in F_0 voorkomen bij "-nacht" en bij "vorst"; beide zijn beklemtoond en deze bewegingen zijn ook na de stilering in fig. 6.2b nog duidelijk terug te vinden. In de overige delen kan het verloop van F_0 sterk worden vereenvoudigd zonder perceptieve gevolgen. Dat de F_0-beweging bij "vorst" perceptief belangrijk is, kan worden gedemonstreerd door de F_0-contour te veran-

deren tot die van fig. 6.2c. Hierin is de beweging verplaatst van "vorst" naar "in", waardoor de zin niet meer als het 'invallen van de vorst' maar als 'vallen van de vorstin' wordt waargenomen.

Toonhoogteaccenten en de daarmee gepaard gaande F_0 bewegingen spelen een belangrijke rol voor de waargenomen intonatie van spraak. Voor het nederlands zijn ze vrijwel volledig in kaart gebracht door 't Hart en Cohen (1973) en 't Hart en Collier (1975). Met de door hen ontwikkelde intonatiegrammatica is een nagenoeg komplete kwantitatieve beschrijving beschikbaar gekomen van de melodische struktuur van het nederlands. Daarmee kan o.m. synthetische spraak worden voorzien van een perceptief correcte F_0-contour, althans wanneer de ligging van accenten en grammaticale grenzen worden aangegeven, hetgeen vooralsnog met de hand moet gebeuren. Wanneer deze gegevens b.v. in de vorm van framenummers worden ingevoerd kan hieruit dan de F_0-contour worden berekend en aangebracht in een bestaande A/P-file. Een programma waarmee dit kan worden uitgevoerd is RNK (Zelle ea, 1983). In fig. 6.2d zien we een voorbeeld van zo'n kunstmatige toonhoogtecontour die met RNK is aangebracht. Deze verschilt wel hoorbaar van de oorspronkelijke contour uit fig. 6.2a, maar klinkt toch volkomen natuurlijk, met de klemtoon op de juiste plaatsen, en zou door de spreker ook zo kunnnen zijn uitgesproken.

Hoewel de uitkomsten van dit intonatieonderzoek voor het nederlands niet zonder meer toepasbaar zijn op andere talen is de methode, met gebruik van het analyse-resynthesesysteem, dat wel. Zo zijn ook voor het brits engels perceptief belangrijke F_0-bewegingen opgespoord via analyse, stilering en resynthese (de Pijper, 1983; N.J.Willems, 1982) en wordt nu intensief gewerkt aan een intonatiegrammatica voor het brits engels (N.J.Willems, 1982).

We hebben hiermee in het kort aangegeven hoe het systeem voor analyse en resyntheses wordt toegepast bij het intonatieonderzoek door de F_0-contour als het ware zo zuinig mogelijk te beschrijven, de perceptief onbelangrijke details weg te laten en mede daardoor synthetische spraak van een eenvoudige en bevredigende intonatie te kunnen voorzien. In de volgende paragraaf zullen we laten zien hoe ook op de andere parameters bezuinigd kan worden.

6.2. TOEPASSING BIJ OPSLAG EN REPRODUKTIE VAN SPRAAK

6.2.1. Zuinige codering van spraak

Industriële toepassing van het analyse-resynthese-systeem ligt op het terrein van zuinige codering van spraak voor uitgifte door appa-

raten. Gewone digitalisering van de golfvorm met 10 kHz samplefrekwentie en een nauwkeurigheid van 12 bits vereist voor iedere seconde spraak 120 kbits aan opslagcapaciteit. Door inkrimping van het frekwentiegebied (lagere samplefrekwentie), door beperking van de nauwkeurigheid (minder bits per sample) of door speciale wijzen van coderen (b.v. deltamodulatie) kan de informatiestroom wel beperkt worden maar dat veroorzaakt al snel hoorbare aantasting van het spraaksignaal. Door de spraak te beschrijven met een bron-filter model kunnen we op zich al flink bezuinigen en wordt al veel in de golfvorm aanwezige redundantie verwijderd. Immers de modelparameters veranderen ongeveer met de snelheid waarmee de akoestische eigenschappen van het mondkanaal wijzigen en dat is veel trager dan de golfvorm zelf. Een gewone analyse geeft dan ook al ongeveer een factor 10 besparing tov het oorspronkelijke 120 kb/s PcM-signaal. Bij de gewoonlijk toegepaste frameduur van 10 ms kost opslag van 1 seconde spraak ongeveer 12 kbit, als we met 8 filtercoëfficiënten werken en elke parameter met 12 bits coderen. Hierop kan nog aanzienlijk verder bezuinigd worden door de afzonderlijke parameters met minder bits, dus grover te kwantiseren en door de frameduur verder te vergroten. Daarbij is een reduktie mogelijk met nog eens een factor 10 tot ongeveer 1 kbit/s.

We zien daarvan een voorbeeld in fig. 6.3. Per frame is hier het aantal bits beperkt tot in totaal 32 waarvan er 30 voor de parameters zelf zijn gebruikt en 2 bits om de frameduur te coderen. De bitverdeling over de afzonderlijke parameters zullen we in de volgende paragraaf verder bespreken. Bij deze zuinige codering is gebruik gemaakt van het feit dat kleine afwijkingen in de ligging van resonanties niet of nauwelijks kunnen worden waargenomen en dat ook de bandbreedte weinig kritisch is voor de spraakperceptie (Flanagan 1972). Verder draagt de ligging van de hogere resonanties relatief weinig bij tot de verstaanbaarheid van de spraak en deze mogen daarom ook vrij ruw worden gekwantiseerd (F3) of zelfs helemaal konstant worden gemaakt (F4 en F5). We zien dat geïllustreerd in fig. 6.3a. Bij de hier toegepaste frameduur van 8 ms is de spraak dan gecodeerd met 4 kbit/s. Een verdere informatiereduktie is mogelijk door de frameduur tot bv 32 ms te vergroten. In fig. 6.3c is dat achteraf gebeurd door op iedere 3 frames er 2 weg te laten en hun parameterwaarden lineair te interpoleren. Deze spraak is dus met slechts 1 kbit/s gecodeerd en verschilt niet hoorbaar van de 4 kbit/s versie van fig. 6.3a. In het algemeen is een frameduur van 8 ms ook lang niet overal nodig; er zijn stukken (b.v klinkers) waar zo weinig in het verloop van de parameters verandert dat zonder bezwaar voor de spraakkwaliteit een aanzienlijk langere frameduur zou kunnen worden gekozen.

Een grote en vaste frameduur zoals in fig. 6.3c kan echter niet altijd worden toegepast; hoge spreeksnelheid en veel plofklanken

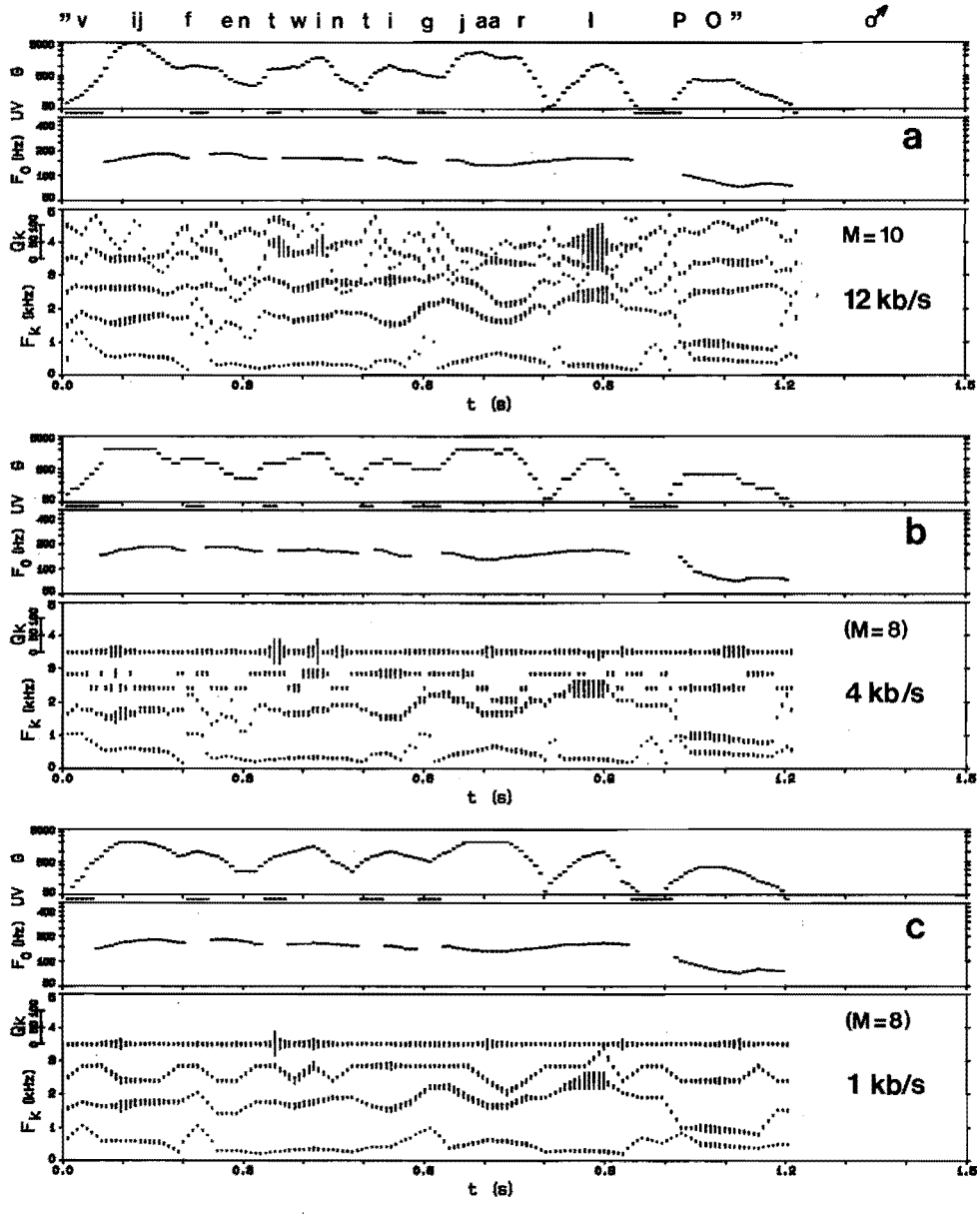

Fig.6.3. Bovenste plaatje (a): standaard analyseresultaten voor een stukje mannenspraak, geanalyseerd met M = 10 filtercoëfficiënten en gecodeerd met ongeveer 12 kbits/s. Bij (b): een zuinige codering met 4 kbit/s, waarbij de parameters van ieder frame (8 ms duur, 8 kHz samplefrekwentie) met in totaal 32 bits zijn gecodeerd. Bij (c): verdere bezuiniging tot 1 kbit/s door de frameduur te vergroten tot 32 ms en tussen de frames de parameters lineair te interpoleren.

kunnen dat uit perceptief oogpunt ongewenst maken. Wel kan een variabele frameduur flinke bezuinigingen opleveren maar we hebben (nog) geen bevredigende algoritmen kunnen vinden om die frameduur automatisch aan te passen aan de snelheid waarmee de parameters veranderen. Dat moet nu nog grafisch interaktief gebeuren, maar daarmee kan dan in de meeste gevallen een gemiddelde frameduur van 32 ms worden bereikt. Een bezuiniging tot minder dan 1 kbit/s is vaak mogelijk zonder de spraak ontoelaatbaar aan te tasten.

Door de coëfficiënten van de 2e-orde resonanties vrij grof te kwantiseren, de bits zorgvuldig over de parameters van het frame te verdelen en de frameduur te variëren kan dus een faktor 100 worden bespaard op de omvang van het geheugen waarin de spraak moet worden bewaard. Daarmee kan dan voldoende spraak in een of enkele chips (thans in grootte variërend van 16 tot 64 kbytes) worden opgeslagen om spraakuitgifte aantrekkelijk te maken voor praktische toepassing in eenvoudige en robuuste digitale systemen.

Deze reduktie gaat momenteel bij sommige stemmen nog ten koste van enig verlies van kwaliteit, maar voor een aantal toepassingen waarbij een voor het systeem geschikte spreker kan worden uitgekozen is die kwaliteit voorlopig voldoende. Er wordt echter nog volop gewerkt aan het vinden van een optimale kwantisering en van verdere bewerkingen van de parameters om de kwaliteit van deze zuinige gecodeerde spraak te verbeteren.

In de volgende paragraaf zullen we bespreken hoe de synthese van deze spraak in 'hardware' wordt uitgevoerd door de 'spraakchip'.

6.2.2. De spraakchip

Enkele jaren geleden is door Willems de aanzet gegeven om de synthese en uitgifte van zuinig gecodeerde spraak in hardware te realiseren. Met deze 'voice response unit' (van Essen en Willems, 1978) kon spraak, gecodeerd in 1 kbit/s en opgeslagen in een prom, ten gehore worden gebracht. De hiermee verkregen resultaten waren dermate veelbelovend dat vervolgens door Philips-Elcoma (van Brück en Teuling, 1982, Bierlaagh, 1982) het gehele syntheseproces is ondergebracht in één enkele chip, de MEA8000, die nu op de markt verkrijgbaar is (Willems en Bierlaagh, 1983).

Deze chip, schematisch weergegeven in fig. 6.4, produceert spraak met behulp van 4 resonantiefilters en bevat behalve het eigenlijke synthesedeel ook een 8-bits digitaal-analoogomzetter alsmede een tabel waarmee de 32 bits van ieder frame worden vertaald in parameterwaarden voor frameduur, bronsignaal en synthesefilter. Tevens voorziet een interface in de koppeling tussen de MEA8000 en de meeste gangbare

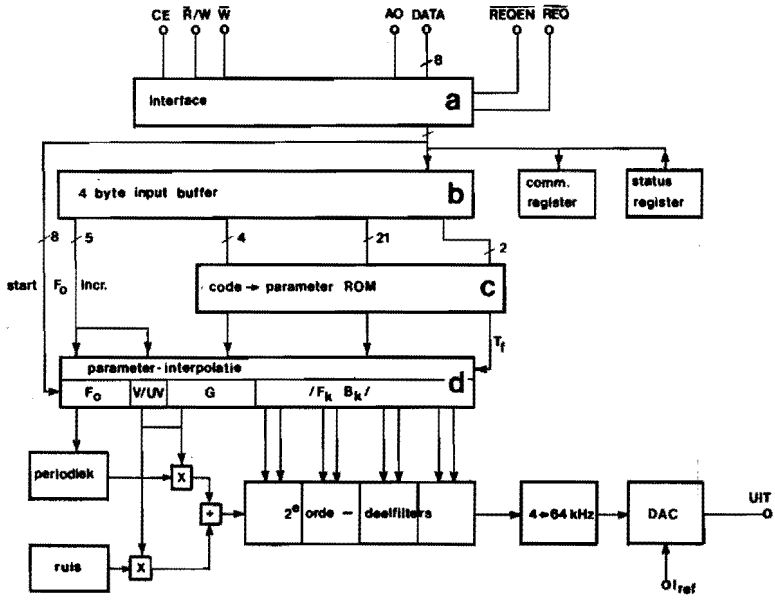

Fig.6.4. Blokschema van de spraakchip MEA8000. Blok (a) verzorgt het (seriële) transport van data uit een (micro)computer of uit de proms waarin de gecodeerde frames liggen opgeslagen. Blok (b) is een buffer waarin de 4 bytes van ieder frame worden verzameld. In blok (c) worden de bitpatronen omgezet naar parameterwaarden die vervolgens in blok (d) worden geïnterpoleerd. Pas dan volgt de eigenlijke synthese met (onderin het plaatje) de cascade van vier 2^e orde deelfilters, geëxciteerd door een periodiek signaal of door ruis.

---

microprocessoren en/of de prom(s) waarin de gecodeerde spraak ligt opgeslagen.

Het eigenlijke synthesedeel is in essentie een hardware implementatie van het bron-filter model zoals we dat in de voorgaande hoofdstukken hebben besproken. Maar op enkele punten wijkt de hardware versie daarvan af.

Allereerst levert de periodiekbron in plaats van een impuls een zaagtandvormige spanning (van Essen en Willems, 1978). De reden daarvoor is dat de piekwaarde van een zaagtandvormig ingangssignaal voor de filters aanzienlijk kleiner is dan van een impuls met dezelfde energie. Op de chip wordt de berekening van de signalen in de filters met integer getallen uitgevoerd en dan treedt bij een zaagtand minder snel oversturing op in de tussenresultaten dan bij een impuls als ingangssignaal. Bij de softwaresynthese treedt zo'n overflow niet op omdat daar met real getallen wordt gerekend die praktisch onbegrensde

waarden kunnen aannemen. Overigens vervalt op de chip door toepassing van een zaagtand de 1e orde de-emfase, die de pre-emfase van de analyse compenseert. Een zaagtand is immers (bij benadering) de de-emfase van een impuls.

Een tweede verschil is dat op de chip de periodiek- en de ruisbron ieder van een eigen amplituderegeling zijn voorzien. Daarmee kunnen in principe dus ook mengvormen van stemhebbende en stemloze klanken worden gerealiseerd. Van deze mogelijkheid wordt echter bij de analyse en codering thans geen gebruik gemaakt omdat algoritmen die tot perceptief bevredigende klanken leiden vooralsnog ontbreken.

Een derde verschil met de softwaresynthese ligt bij het synthesefilter zelf. Op de chip bestaat dat uit een cascade van 2e orde filters. Daarmee hoeven de 2e orde polynomen dus niet, zoals in hfdst 3 is besproken, tot een 8e orde polynoom te worden uitvermenigvuldigd. Maar dat betekent wel dat de coëfficiënten van de afzonderlijke 2e orde secties, de cr-parameters, gesorteerd moeten worden conform par. 2.5. Tevens worden ze daarna begrensd zodat de vier deelfilters altijd resonerend zijn. Dit laatste is wat de chip betreft niet strikt noodzakelijk; de 2e orde synthesesfilters van de MEA8000 kunnen ook reële polen aan. Maar in de tabel (Bierlaagh, 1982) waarmee de 'binnenkomende' frames worden gedecodeerd naar filtercoëfficiënten zijn c en r beperkt tot $|c| < 1$ en $r > 0$. Met iedere cr-kombinatie uit deze tabel is dus steeds een afstemfrekwentie F en bijbehorende bandbreedte B of kwaliteitsfaktor Q geassocieerd.

Het synthesedeel berekent de golfvorm (8 kHz samplefrekwentie) uit de opeenvolgende frames waarvan de parameters in 32 bits zijn gecodeerd (tabel 6.1). Daarvan zijn er 30 voor bron- en filterparameters gebruikt en 2 voor de frameduur, die dus 4 waarden kan hebben: 8, 16, 32 of 64 ms (Willems en Bierlaagh, 1983). De vereiste opslagcapaciteit of bitsnelheid kan dus variëren tussen 4000 en 500 bits per seconde spraak. Per frame is de synthese nog in 8 stapjes onderverdeeld (van Essen en Willems, 1978), zodat bij een frameduur van b.v. 64 ms de parameters om de 8 ms worden bijgesteld volgens een lineaire interpolatie tussen twee opeenvolgende frames.

Voor het bronsignaal zijn 9 bits beschikbaar (tabel 6.1): 4 voor de amplitude en 5 voor de grondtoon die gecodeerd is als verschilfrekwentie t.o.v. F_0 van het vorige frame. Een van de 32 waarden hiervan is gereserveerd als stemloos-aanduiding. Door de verschilfrekwentie te coderen (van Essen en Willems, 1978) worden per frame minstens 2 bits bespaard omdat F_0 zelf meestal relatief langzaam verandert en het coderen daarvan minstens 7 bits per frame zou vereisen om duidelijk hoorbare stapjes in de waargenomen toonhoogte te vermijden. Alleen het eerste frame van een uiting bevat de waarde van F_0 zelf, gecodeerd in

naam	bits	parameter MEA8000
T_f	2	frameduur (8, 16, 32 of 64 ms)
ΔF_o	5	toonhoogte-increment cq UV
G	4	amplitude-versterkingsfaktor
F_1	5	afstemfrekwentie 1e resonantie
F_2	5	afstemfrekwentie 2e resonantie
F_3	3	afstemfrekwentie 3e resonantie
F_4	0	afstemfrekwentie 4e resonantie
B_1	2	bandbreedte 1e resonantie
B_2	2	bandbreedte 2e resonantie
B_3	2	bandbreedte 3e resonantie
B_4	2	bandbreedte 4e resonantie

Tab.6.1. Verdeling van de 32 bits per frame over de afzonderlijke parameters. De frameduur is variabel en kan 4 waarden aannemen. De grondtoonfrekwentie F_o is in 5 bits gecodeerd als het verschil met het vorige frame; van de 32 mogelijke waarden is er een gereserveerd als stemloosaanduiding. Van het 4e deelfilter is de afstemfrekwentie F_4 constant 3500 Hz (0 bits). Omdat de chip met een samplefrekwentie van 8 kHz werkt is er geen 5e deelfilter.

8 bits.

Voor de filterparameters zijn 21 bits beschikbaar. Daarvan zijn er 2 voor elke bandbreedte, 5 voor F_1 en F_2 en 3 voor F_3. F_4 heeft een vaste waarde van 3500 Hz (0 bits). Het resogram van aldus gekwantiseerde spraak is in de al eerder gepresenteerde fig. 6.3b en c weergegeven. We zien dat de contouren van de lage resonanties F_1 en F_2 nauwelijks zichtbaar verschillen van de ongekwantiseerde versie in fig. 6.3a. In het algemeen klinkt de gekwantiseerde spraak ook redelijk tot goed; soms zijn er nauwelijks verschillen te horen met de ongekwantiseerde versie, maar dat hangt sterk af van de spreekstem. Voor sprekers of spreeksters waarbij relatief veel reële nulpuntenparen in de oorspronkelijke analyseresultaten aanwezig zijn levert de huidige chip duidelijk nog geen goede resultaten. De oorzaak hiervan is waarschijnlijk dat de ordening van de 2e orde filterparameters dan geen optimale resultaten oplevert. Voor die stemmen is het aantal filtercoëfficiënten waarmee wordt geanalyseerd vrij kritisch en kan het voorkomen dat een analyse met 6 in plaats van 8 coëfficiënten betere resultaten levert. De huidige chiptabel is daar echter nog niet op berekend en aan een verbetering daarvan wordt gewerkt.

Bij het onderzoek naar een, perceptief zo goed mogelijke, zuinige codering van spraak speelt het analyse-resynthese-systeem uiteraard een belangrijke rol. Kwantisering en codering zoals hierboven is beschreven kan automatisch worden verricht met b.v. het programma KWC. Hiermee worden gewone P-files met analysedata gekwantiseerd en de

gecodeerde frames vervolgens opgeslagen in een zgn chip-file, die dan via een terminallijn vanuit de VAX naar de MEA8000 wordt getransporteerd en verklankt. Om de chip via zo'n gewone terminallijn te kunnen besturen is door Polstra (1981) een interface ontwikkeld waarin onder meer een extra buffer is aangebracht om de bij time sharing optredende variabele wachttijden ook bij de hoogste bitsnelheid van 4 kbit/s probleemloos te overbruggen. Net als bij de uitgifte van N-files met de digitale golfvorm, heeft de gebruiker de beschikking over commando's waarmee een of meer chip-files direkt via de spraakchip worden verklankt. Ook in interaktieve programma's zoals CHF kan een willekeurig spraaksegment door indrukken van een toets gecodeerd worden met 4, 2, 1 of 0.5 kbit/s waarna de chip de hardwaresynthese, digitaal-analoogconversie en uitgifte voor zijn rekening neemt. Daarmee kan dus ook direct worden beluisterd welke gevolgen de kwantisering heeft voor de spraak, hoe ver we mogen gaan bij het vergroten van de frameduur en welke sprekers goede of minder goed klinkende spraak opleveren.

Behalve vanuit de VAX kan de MEA8000 ook heel goed vanuit een microcomputer worden bedreven (Bierlaagh, 1982). Programma's om files met gecodeerde spraak van b.v. de VAX over te hevelen naar floppy discs van b.v. een Apple of Exorciser zijn beschikbaar.

Tot slot van deze paragraaf willen we nog de toepassing van de spraakchip voor de praktijk noemen, waarbij de MEA8000 de gecodeerde data betrekt uit een of meer proms (Bierlaagh, 1982). In de handel zijn nu op ruime schaal proms verkrijgbaar met een geheugencapaciteit van 64 kbyte of 512 kbit, overeenkomend met 16000 frames van 32 bits. Afhankelijk van de toegepaste frameduur is dat minimaal 128 seconden tot maximaal ruim 17 minuten spraak. Ook voor het vullen van zo'n prom zijn commando's beschikbaar, waarmee de geanalyseerde, eventueel F_0- en VUV-gecorrigeerde en gecodeerde files via een floppy disc vanuit een Apple of Exorciser in een prom worden ingelezen, samen met een tabel met beginadressen van de opeenvolgende woorden of zinnen (Bierlaagh, 1982). Alle uitingen zijn dus op afroep onmiddellijk hoorbaar te maken. Daarmee zijn talloze praktische toepassingen te realiseren voor situaties waarin geschreven overdracht van informatie onnodig of ongewenst is, zoals bij omroep van snel wiselende of verouderende gegevens, of gesproken instrukties of waarschuwingen. Ook voor situaties waarin visuele communicatie niet mogelijk is zijn toepassingen te noemen. Zo is door Kroon en de Braal (1980) voor visueel gehandicapten een spellende typemachine, de Typofoon ontwikkeld. Een ander voorbeeld is een apparaatje waarmee diabetici zelf thuis hun bloedsuikergehalte kunnen bepalen, dat door Waterham (1983) is aangepast en voorzien van spraakoutput zodat ook blinden of slechtzienden kennis kunnen nemen van het meetresultaat.

Grote voordelen van deze chips zijn dat woorden en (delen van) zinnen zonder tijdverlies in willekeurige volgorde kunnen worden uitgesproken, dat de apparatuur afgezien van een luidspreker geen bewegende (en slijtende) onderdelen bevat en dat ze veel kleiner en lichter kan zijn dan conventionele apparatuur om spraak te reproduceren.

Een essentiele beperking is uiteraard dat alleen de woorden en zinnen kunnen worden weergegeven die ooit eerder door een menselijke stem zijn ingesproken. Hoe het analyse-resynthese-systeem wordt toegepast bij de ontwikkeling van een veel krachtiger methode van spraakuitgifte zullen we in de volgende paragraaf behandelen.

6.3. TOEPASSING BIJ SAMENSTELLEN VAN NIEUWE SPRAAK

We hebben in de vorige paragraaf gezien hoe de spraakuitgifte van het daar besproken systeem beperkt is tot reproduktie van eerder door een menselijke stem voortgebrachte spraak. Weliswaar kan enige verandering in de samengestelde uiting worden aangebracht, b.v. door bij de uitgifte de volgorde van woorden, zinsdelen of zinnen te wijzigen maar daarbij moeten vaak ruime pauzes tussen de woorden worden aangebracht wil de spraak goed verstaanbaar blijven. Erg flexibel is zo'n systeem niet; nieuwe woorden en zinnen moeten steeds als zodanig uitgesproken, geanalyseerd, gekodeerd en opgeslagen worden.

Veel verderreikende mogelijkheden zou een systeem bieden dat, naar analogie van b.v. tekst op een beeldscherm, willekeurige uitingen samenstelt door elementaire klanken, uit een begrensde en liefst niet te grote verzameling, aan elkaar te rijgen. Bij de vraag welke eenheden we daarbij als elementaire klanken moeten kiezen kunnen we aan twee uitersten denken: enerzijds woorden (of eventueel groepen woorden) en anderzijds fonemen, de kleinste betekenisdragende spraakklanken.

Woorden hebben als praktisch bezwaar hun grote aantal. Bovendien levert het zonder extra pauzes aaneenrijgen van los uitgesproken woorden geen goed verstaanbare spraak en is van een vloeiend verloop meestal geen sprake. Natuurlijke, aaneengesloten uitgesproken spraak verschilt zowel perceptief als fysisch van aaneengeschakelde, los uitgesproken woorden. Woordgrenzen zijn in een zin nauwelijks of niet fysisch vast te stellen en ook de woordduur is in zinsverband meestal aanzienlijk korter dan los uitgesproken. We zien daarvan een voorbeeld in fig. 6.5 waarin de zin "de temperatuur van het koelwater is te hoog" is weergegeven, samen met dezelfde zin maar dan samengesteld uit los uitgesproken woorden die in de juiste volgorde aaneen zijn geregen. Niet alleen de duuropbouw, ook het verloop van afzonderlijke

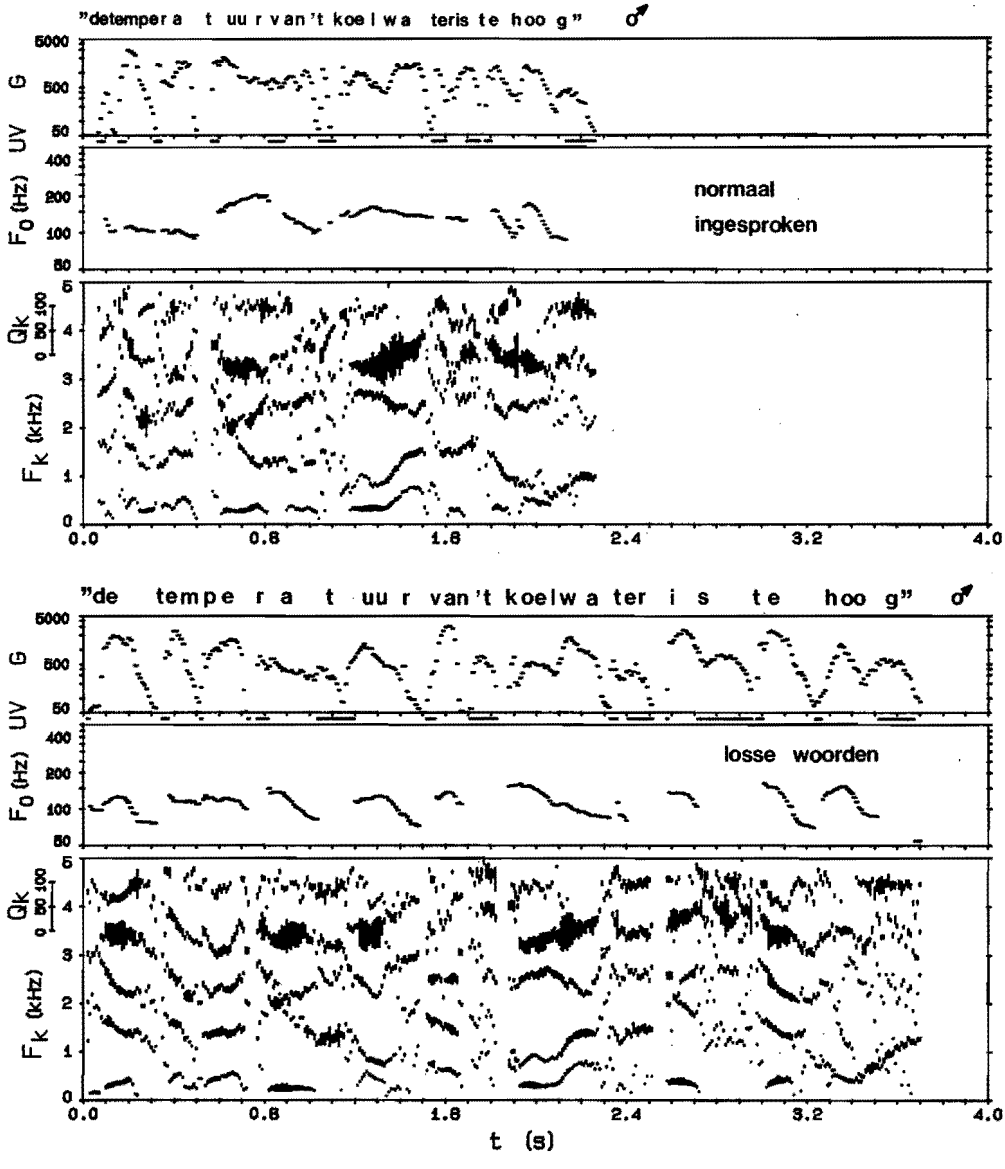

Fig.6.5 Analyseresultaten voor de zin "de temperatuur van het koelwater is te hoog" (mannenstem, M = 10). Boven: normaal ingesproken. Onder: samengesteld door losse woorden, die eerst in omgekeerde volgorde zijn ingesproken, in de juiste volgorde aan elkaar te rijgen en dan te analyseren.

parameters zoals de toonhoogte F_o en (in minder mate) de ampitude G vertoont grote verschillen. Om tot vloeiende spraak te komen zullen deze dus (volgens regels) moeten worden gemanipuleerd.

Wanneer we in plaats van woorden fonemen als elementaire 'bouwstenen' zouden nemen heeft dat weliswaar het voordeel dat we er daarvan (voor het nederlands) niet meer dan ongeveer 60 hoeven op te slaan, maar ook dit levert na zonder meer aaneenrijgen geen verstaanbare, laat staan natuurlijk klinkende spraak. Fonemen zijn evenmin als woorden vormvast; de fysische realisatie hangt in lopende spraak sterk af van voorafgaande en volgende fonemen. Vooral aan het energiespectrum is dat duidelijk zichtbaar. Fig. 6.6. geeft daarvan een voorbeeld. Het spectrum van een stukje ch uit "schoon" heeft een heel andere samenstelling dan dat van de ch uit "schip", door de verschillende klinkers oo en i die volgen op de ch. Hier zien we geïllustreerd hoe een en hetzelfde foneem door sterk verschillende spraaksignalen wordt gerealiseerd, terwijl we toch beide als ch waarnemen. Verder zijn vooral de overgangen tussen de fonemen vaak perceptief erg belangrijk. Juist op die overgangen variëren de fysische eigenschappen van het spraaksignaal sterk en zou een groot aantal regels nodig zijn om de parameters daar zo te wijzigen dat verstaanbare en redelijk vloeiende spraak onstaat. Zo'n systeem van synthese door (heel veel) regels is voor het amerikaans engels ontwikkeld door Klatt (1976), en voor het nederlands door Slis et al (1977).

Een derde mogelijkheid is om een tussenweg te kiezen. Door Elsendoorn wordt thans, in navolging van Olive (1977-1979), gebruikt gemaakt van zgn. difonen als bouwstenen (Elsendoorn en 't Hart,1983). Difonen zijn stukjes spraakklank rond de overgang tussen twee fonemen, ruwweg van halverwege het ene tot halverwege het volgende foneem. Daarmee wordt die perceptief zo belangrijke overgang in z'n geheel

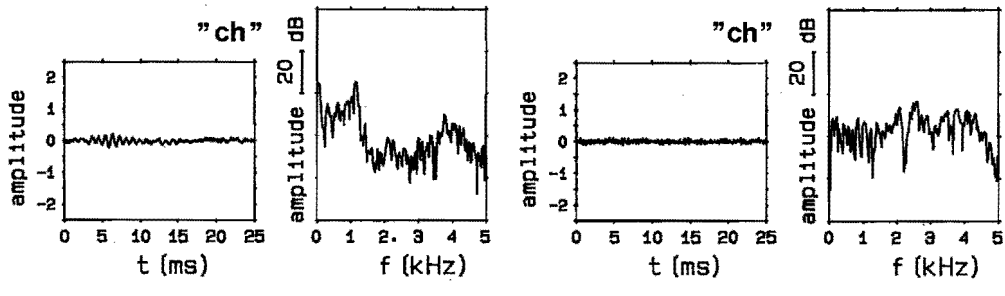

Fig.6.6. Voorbeelden van twee realisaties van de ch, links uit "schoon" en rechts uit "schip".

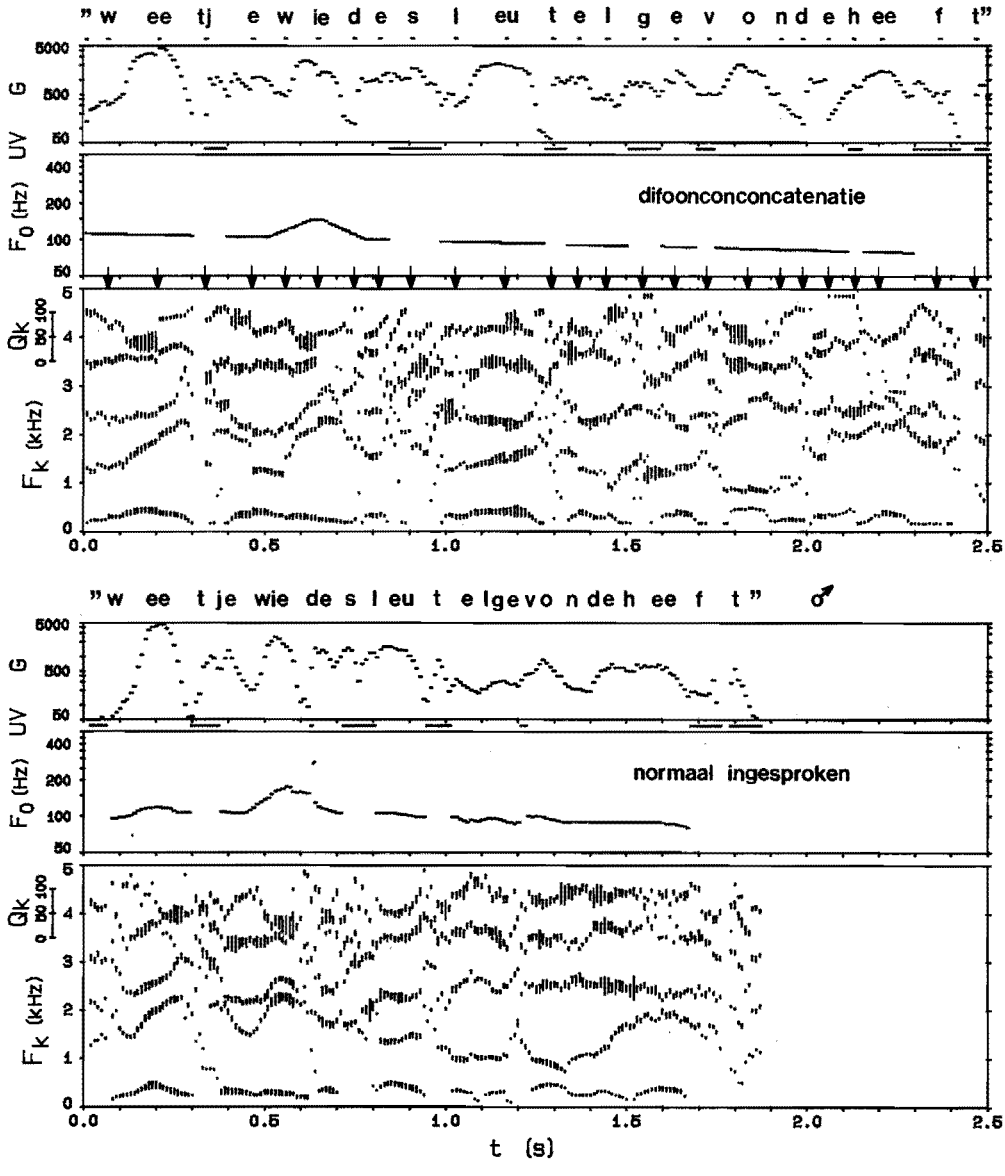

Fig.6.7. Boven: difoonconcatenatie van de zin "weet je wie de sleutel gevonden heeft", samengesteld door de parameters van de afzonderlijke difonen aan elkaar te rijgen en de file daarna te voorzien van een toonhoogtecontour. Onder: dezelfde zin door dezelfde spreker in zijn geheel ingesproken en geanalyseerd.

(in geanalyseerde vorm) opgeslagen en komen de lassen op dat deel van de klanken te liggen waar de fysische eigenschappen het minst snel veranderen.

In principe moeten nu zoveel difonen worden opgeslagen als er foneemkombinaties zijn. Dat zijn er geen 3600 omdat in het nederlands lang niet alle kombinaties voor komen. In de praktijk blijken ongeveer 1400 stuks voldoende voor een verzameling waaruit willekeurige uitingen kunnen worden gevormd. Gebleken is dat het zonder meer aaneenrijgen van difonen al vaak redelijke en soms zelfs verrassend goed verstaanbare spraak oplevert. Dat deze werkwijze nog zonder regels voor het aaneenrijgen al zo'n goed resultaat oplevert komt waarschijnlijk doordat de perceptief belangrijke foneemovergangen nu in hun geheel bewaard zijn gebleven. Wel laat de mate van natuurlijkheid nog te wensen over, in hoofdzaak omdat de tijdsopbouw ('ritmiek') van de uiting nog niet correct is. Niet alleen is de totale lengte van uit difonen opgebouwde zinnen meestal te groot, ook de nog onjuiste duuropbouw voor de klinkers en medeklinkers onderling maakt dat de zin nog niet altijd vloeiend loopt.

Ook aan dit probleem wordt momenteel intensief onderzoek verricht en het zal duidelijk zijn dat ook hierbij het analyse-resynthesesysteem een belangrijke rol speelt. De difonen zelf zijn ermee gemaakt, door afsplitsing uit de analyseresultaten van bestaande dan wel nonsenswoordjes waarin de betreffende foneemovergangen voorkomen. Alle difonen zijn opgeslagen in korte P-files, ter lengte van 10 à 20 frames en zijn dus direkt toegankelijk om tot een langere file aaneen te rijgen. Daarna wordt op de zo gevormde P-file de gewenste F_0-contour aangebracht, b.v met het programma RNK (Zelle ea, 1983) en kan de synthese beluisterd en beoordeeld worden.

Een voorbeeld van zo'n difoongeconcateneerde zin is in fig. 6.7a weergegeven. Hoewel in dit resogram de naden tussen de afzonderlijke difonen vaak duidelijk zichtbaar zijn in zowel de amplitude als de FQ-parameters, blijken deze toch slechts in enkele gevallen aanleiding te geven tot hoorbare discontinuïteiten. Ter vergelijking is in fig. 6.7b het resogram gegeven van dezelfde zin, maar nu door dezelfde spreker als een geheel ingesproken. Duidelijk is te zien hoe de totale duur na concatenatie veel langer is en hoe ook de afzonderlijke woorddelen verschillen qua duuropbouw. Eerder in dit hoofstuk hebben we aangegeven dat met het analyse-resynthesesysteem ook de duur van ieder gewenst spraaksegment naar wens veranderd kan worden, door de lokale frameduur te wijzigen. De hoop is dan ook dat langs deze weg de regels kunnen worden opgespoord volgens welke de duren van de afzonderlijke difoon(delen) moeten worden bijgesteld, om zo in de toekomst tot automatische samenstelling van spraakuitingen te komen waarvan ook de tijdsopbouw perceptief correct is.

7 N A B E S C H O U W I N G

We hebben in de vorige hoofdstukken principe, uitvoering en enige toepassingen beschreven van een computersysteem voor analyse en resynthese van spraak in termen van een klein aantal fysische parameters: grondtoonfrekwentie F_0, amplitude G, stem/stemloosparameter VUV en de coefficienten van een vijftal 2e orde filters die de spectrale omhullende karakteriseren. De parameters van dit bron-filtermodel sluiten nauw aan bij kenmerken in de spraakperceptie zoals: luidheid, stemhebbend/stemloosheid, toonhoogte en timbre. We hebben ook laten zien hoe we op basis van dit model synthetische spraak kunnen genereren die in het algemeen voor lopende spraak (woorden, zinnen), perceptief goed overeenkomt met de oorspronkelijke spraak. Bij sommige stemmen is die overeenkomst zo groot dat de resynthese nauwelijks van het origineel is te onderscheiden. Soms echter zijn er duidelijke verschillen waar te nemen en klinkt de resynthese, minder transparant of 'nasaler' dan het origineel. Dit blijkt vooral af te hangen van de spreekstem; de ene spreker voldoet kennelijk beter aan het model dan de andere. Daarnaast hebben we in hfdst 5 gezien dat bij losse (nonsens) woordjes, gevormd door klinker-medeklinkercombinaties, een aantal medeklinkers na resynthese duidelijk slechter wordt herkend dan in de originele versie. Op beide terreinen zal onderzoek nodig zijn om ook voor moeilijke stemmen de perceptieve overeenkomst tussen input en output van het systeem verder te vergroten.

Het analyse-resynthesesysteem is op de eerste plaats ontwikkeld als gereedschap voor het spraakonderzoek. Perceptief belangrijke fysische eigenschappen in het spraaksignaal kunnen hiermee op snelle, flexibele en interaktieve wijze worden gemanipuleerd. We hebben bij de keuze van de parameters waarin het spraaksignaal wordt ontleed geëist dat ze aansluiten bij relevante kenmerken in de spraakperceptie. Voor de bronparameters is die aansluiting duidelijk. De filterparameters worden in ons systeem weergegeven in de vorm van 2e-orde resonanties die direct gerelateerd kunnen worden aan formanten, de in de fonetiek gebruikelijke karakterisering voor de omhullende van het korte-termijn energiespectrum van spraak. Representatie van de analyse-resultaten in de vorm van een resogram, waarin de FQ-parameters van deze resonanties zijn weergegeven als funktie van de tijd, sluit ook aan bij de traditionele weergave van het spectrum door een spectrogram ('sonagram'). Het resogram is daarvan een vereenvoudigde afspiegeling.

De praktijk heeft inmiddels aangetoond dat ons systeem voor analyse en resynthese flexibel toepasbaar is in het spraakonderzoek. Binnen het IPO wordt het intensief gebruikt op diverse terreinen zoals het onderzoek aan intonatie, klemtoon, prosodie en concatenatie van spraak. Spraakinname in de computer, analyse, resynthese en uitgifte

kunnen voor een vrijwel onbeperkt aantal files practisch automatisch worden uitgevoerd zonder veel training of speciale kennis. Het beschikbaar zijn van faciliteiten voor interaktief experimenteren met spraakparameters heeft ertoe bijgedragen dat het systeem een centrale plaats in het spraakonderzoek op het IPO heeft verworven. Ook buiten het IPO wordt gebruik gemaakt van ons systeem, o.a. bij het analyseren van klinkers en het samenstellen van stimuli voor experimenten met spraakaudiometrie. Daarnaast is de komplete software voor analyse, kwantisering en zuinige codering bij diverse vestigingen van Philips in binnen- en buitenland in gebruik voor spraakuitgifte via de MEA8000 spraakchip.

In de inleiding hebben we ook als eis gesteld dat de parameters waarin de spraak wordt geanalyseerd rechtstreeks en automatisch uit het spraaksignaal zelf moeten zijn te bepalen. Wat dit laatste betreft worden de grondtoonfrekwentie F_0 en de stem/stemloosparameter VUV nu nog niet altijd automatisch foutloos berekend. Moeilijke stemmen, bandfiltering (telefoonspraak), nagalm of ruis in de originele spraak kunnen ertoe leiden dat de automatisch verkregen resultaten achteraf verbeterd moeten worden teneinde na resynthese een betere perceptieve overeenkomst met het origineel te verkrijgen. Weliswaar is die correctie snel en handig grafisch interaktief uit te voeren, maar zowel automatische bepaling van de toonhoogte als de stem/stemloosparameter kunnen o.i. door toekomstig onderzoek beslist nog verbeterd worden.

Datzelfde geldt voor de toepassing van het systeem voor zuinige codering van spraak. Op dit terrein zijn reeds veelbelovende resultaten geboekt. Met de MEA8000 is het mogelijk gebleken om redelijk klinkende spraak te produceren die met minder dan 1 kbit/s is gecodeerd. Maar ook hier geldt dat de bereikte kwaliteit vrij sterk sprekerafhankelijk is. In het algemeen is gebleken dat bij die stemmen waarvan de analyse relatief veel reële nulpuntenparen oplevert ook vaak problemen optreden. De zuinige codering is vrij gevoelig voor een analyse met het 'juiste' aantal filtercoëfficiënten. Zo kan bij vrouwenstemmen b.v. een analyse met M = 6 nodig zijn, terwijl de resynthese in de huidige chip met M = 8 werkt. Verder onderzoek op dit gebied vindt thans plaats.

Ook onderzoek met een meer ingewikkeld samengesteld brongeluid dan de nu gebruikte enkele impuls per grondtoonperiode of ruis, om zo tot betere resynthese van zuinig gecodeerde spraak te komen, is thans in volle gang.

De eindconclusie is dat het ontwikkelde systeem, hoewel zeker nog voor verbeteringen vatbaar, zich een duidelijke plaats heeft verworven bij het fundamenteel spraakonderzoek in het algemeen en bij de ontwikkeling van sprekende apparaten in het bijzonder.

SAMENVATTING

Het hier beschreven onderzoek betreft de ontwikkeling van een software computersysteem voor analyse en resynthese van natuurlijke spraakuitingen in termen van een klein aantal perceptief relevante parameters. Dit analyse-resynthesesysteem vormt een belangrijk hulpmiddel bij het experimenteel spraakonderzoek omdat hiermee een aantal fysische parameters van het spraakgeluid op snelle en interaktieve wijze onafhankelijk van elkaar kunnen worden gemanipuleerd. Daarmee kan het belang van deze parameters voor de spraakperceptie worden vastgesteld. Door in het tijdsverloop van de parameters perceptief onbelangrijke details weg te laten kan tevens de struktuur van de gegenereerde (synthetische) spraak sterk worden vereenvoudigd. De hiermee gepaard gaande zuinige beschrijving van het spraaksignaal maakt het mogelijk om zowel opslag als resynthese onder te brengen in enkele chips, voor toepassing in kleine en robuuste apparatuur voor uitgifte van synthetische spraak.

Hoofdstuk 1 beschrijft in het kort de fysica van de menselijke spraakproduktie alsmede een eenvoudig lineair model daarvoor. Dit model, het bron-filtermodel van Fant (1960), is algemeen aanvaard in de experimentele fonetiek en staat centraal in ons systeem. Het beschrijft spraakgeluid met een variabel bronsignaal (periodiek of ruisig) als excitatie voor een variabel filter dat de akoestische eigenschappen van keel-, mond-, neusholtes en lippen weerspiegelt. Het bronsignaal benadert het geluid dat ontstaat door trilling van de stembanden (stemhebbende klanken) of door luchtturbulenties bij een plaatselijke vernauwing in het mondkanaal (stemloze klanken). Dit bronsignaal wordt als funktie van de tijd gespecificeerd met drie parameters: de amplitude G, ruisig of periodiek (VUV) en (indien periodiek) de periode van de grondtoon F_0. Deze parameters zijn nauw gerelateerd aan de perceptieve kenmerken luidheid, stemhebbend/stemloosheid en toonhoogte. Het variabel filter bestaat in ons systeem uit een cascade van tweede-orde deelfilters en wordt als funktie van de tijd weergegeven met een vijftal resonanties die elk gespecificeerd worden met een afstemfrekwentie F en een bandbreedte B of kwaliteitsfaktor Q. Deze resonanties zijn nauw gerelateerd aan formanten, de in de fonetiek veel gebruikte karakterisering van perceptief belangrijke gebieden met relatief hoge energie in het spectrum.

Hoofdstuk 2 behandelt hoe we de parameters van dit model uit het spraaksignaal zelf kunnen berekenen. Bron- en filterparameters van het model variëren relatief langzaam in de tijd en bepalen de fijnstruktuur en de omhullende van het energiespectrum. Omdat het bronsignaal van het model bestaat uit een reeks impulsen of ongecorreleerde ruis

en dus een wit spectrum heeft met vlakke omhullende, kunnen we de parameters van het filter bepalen volgens de techniek van invers filteren of lineaire prediktie. Daarmee berekenen we de coëfficiënten van een hogere-orde analysefilter dat de totale energie minimaliseert en daarmee, bij gegeven orde M, de spectrale omhullende van het spraaksignaal zo vlak mogelijk maakt. Ook de spraak/stiltebeslissing en de bronparameters worden uit het spraaksignaal zelf bepaald. Vervolgens leiden we uit de berekende a-parameters van het M^e orde filter de pq-coëfficiënten af van een cascade van M/2 tweede-orde deelfilters en transformeren deze naar parameters van louter resonerende deelfilters, de FB- of FQ-parameters. Deze FQ-parameters van het analysefilter worden weergeven in de vorm van 'resogrammen', die enigszins zijn te vergelijken met in de fonetiek veel gebruikte spectrogrammen.

Hoofdstuk 3 beschrijft de resynthese. Omdat in het produktiemodel het bronsignaal een vlak, wit spectrum heeft kan het oorspronkelijke spraaksignaal benaderd worden door het analysefilter te inverteren en dit synthesefilter te exciteren met een spectraal vlak signaal: een impulsreeks of ongecorreleerde ruis. Bij gegeven orde M benadert dit synthesefilter de omhullende van het oorspronkelijke spraaksignaal dan zo goed mogelijk. Excitatie van dit synthesefilter met periodieke impulsen of witte ruis levert dan het geresynthetiseerde spraaksignaal.

Hoofdstuk 4 geeft in een fysische evaluatie de beperkingen weer van het analyse-resynthesesysteem. Allereerst beschrijft het model in principe geen stemhebbende wrijfklanken omdat het geen kombinatie van periodiek en ruisig brongeluid kent; het synthesefilter wordt slechts geëxciteerd met impulsen of met witte ruis. Ook plofklanken met hun relatief snelle overgangsverschijnselen worden in principe aangetast omdat de analyse een stationair signaal veronderstelt. Tenslotte worden steile hellingen in de spectrale omhullende aangetast vanwege het beperkte aantal filterparameters. Door deze beperkingen veroorzaakt het systeem fysische verschillen tussen de oorspronkelijke en de geresynthetiseerde spraak.

Hoofdstuk 5 bevat vervolgens een perceptieve evaluatie van het analyse-resynthesesysteem. Voor één spreker is de herkenning van 18 nederlandse consonanten getest met korte (deels nonsens) woordjes. Vergeleken met de oorspronkelijke versie neemt de correctscore na resynthese, gemiddeld over alle consonanten, af van 98% naar 82%. Aangetast worden vooral de stemhebbende plofklank b, de stemhebbende wrijfklanken v en z, de nasalen m, n en ng en de approximanten w en h. Deze aantastingen zijn hoofdzakelijk terug te voeren tot de eerdergenoemde modelbeperkingen die leiden tot fouten bij de automatische stem/stemloosbepaling en de spraak/stiltebeslissing. Ze treden voornamelijk op bij stemhebbende wrijfklanken, waarin periodieke en ruisige componenten van dezelfde grootteorde zijn, en bij stemhebbende plof-

klanken waarin de ruiscomponent van relatief korte duur is.

Hoofdstuk 6 geeft vervolgens een aantal voorbeelden van toepassing van het analyse-resynthesesysteem bij het experimentele spraakonderzoek, zowel in het IPO als daarbuiten. Belangrijk kenmmerk van ons systeem is dat manipulaties en vereenvoudigingen in het verloop van de parameters snel en interaktief zijn uit te voeren, waarbij de perceptieve gevolgen van deze ingrepen direct beluisterd en beoordeeld kunnen worden via de geresynthetiseerde spraak. Ook voor zuinige codering van spraakgeluid, voor uitgifte via de commerciëel verkrijgbare MEA8000 spraakchip, wordt het systeem intensief gebruikt.

Hoofdstuk 7 reikt tenslotte enkele mogelijkheden aan voor verder onderzoek teneinde de met het systeem gegenereerde spraak verder te verbeteren.

SUMMARY

This thesis describes the development of a software computer system for analysis and resynthesis of natural human speech in terms of a small number of perceptually relevant parameters. This analysis-resynthesis system forms a powerful tool for experimental speech research since a number of physical parameters of the speech sound can be manipulated interactively to study their role in speech perception. The structure of the generated synthetic speech can be simplified by deleting those details in the time course of parameters which turn out to be perceptually irrelevant. The resulting economic coding of the speech signal allows storage and resynthesis of speech in small and robust devices for synthetic speech output.

Chapter 1 briefly describes the principles of the physics of human speech production, along with a simple linear speech production model. This model, the source-filter model of Fant (1960), is generally accepted in phonetics and plays a central role in our system. It describes speech sounds as a variable source signal (periodic or noise), excit ing a variable filter that reflects the resonant properties of the human vocal tract. The source signal approximates the sound produced by either the vibration of the vocal folds (voiced sounds) or a turbulent air flow at a local constriction in the vocal tract (unvoiced sounds). The source signal is specified as a function of time with three parameters: amplitude gain G, noise or periodic VUV and, if periodic, the fundamental frequency F_0. These parameters are closely related to the perceptual features: loudness, voiced/unvoiced and pitch. In our system the variable filter consists of a cascade of second order filters, specified as a function of time by 5 resonances, each of which is determined by its tuning frequency F and its bandwidth B or corresponding quality factor Q. These resonances are closely related to formants which are frequently used in experimental phonetics to characterize perceptually important regions of high energy in the spectrum.

Chapter 2 treats the calculation of model parameters from the speech signal itself. Source and filter parameters vary relatively slowly with time and determine the fine structure and envelope of the energy spectrum. Because the source signal consists of a sequence of impulses or uncorrelated noise, both of which have a white spectrum with a flat envelope, the filter parameters can be determined with the technique of inverse filtering or linear prediction. The coefficients of a higher order analysis filter are calculated by minimizing the total energy of the output signal of the filter. For a given linear filter of order M the spectral envelope at the output is then as flat as

possible. The speech/silence decision and the source parameters are also determined from the speech signal itself. From the calculated a-coefficients of the M^{th} order filter, pq-coefficients are determined for a cascade of M/2 second order filters. These pq-coefficients are then transformed into parameters of filters which are forced to be resonant, the FB- or FQ-parameters. The FQ-parameters of the analysis filter are displayed in so called 'resograms' which are, to some extend, comparable with spectrograms, frequently used in phonetics.

Chapter 3 decribes the resynthesis part of the system. Because in the production model the source signal has a flat, white energy spectrum, the original speech signal can be approximated by inverting the analysis filter and exciting this synthesis filter with a spectrally flat signal: a sequence of impulses or uncorrelated noise. For a given order M, this synthesis filter then describes the envelope of the original speech spectrum as well as possible. Excitation of the synthesis filter with periodic impulses or white noise then yields the resynthesized speech signal.

Chapter 4 contains a physical evaluation of the analysis-synthesis system and deals its limitations. Firstly the model is in principle unable to describe voiced fricatives, because it has no combination of periodic and noisy source signals. Secondly plosives, with relatively fast transitions, will be affected, because the signal inside the analysis window is assumed to be stationary. Finally, steep slopes in the spectral envelope of the input signal will be influenced because of the limited number of filter parameters. Owing to these limitations the system results in physical differences between the original and the resynthesized speech signal.

Chapter 5 contains a perceptual evaluation of the system. For one speaker the recognition of 18 dutch consonants has been tested with short (mainly nonsense) words. Compared with the input version of the system the mean correct score of the resynthesized consonants decreases from 98% to 82%. The system mainly affects the plosive b, the voiced fricatives v and z, the nasals m, n and ng and the approximants w and h. The decrease in recognition can be attributed to the model limitations, which cause errors in the automatic voiced/unvoiced and speech/silence decision. These errors mainly occur in voiced fricatives, when periodic and noisy components are of the same order of magnitude, and in plosives, when the noisy component has a relatively short duration.

Chapter 6 gives some examples of application of the analysis synthesis system in experimental speech research. An important property of our system is that stylizations in the time course of the parameters can be effected interactively, easily and quickly. The researcher can immediately listen to the effect of these alterations upon the speech

output and assess their consequences for speech perception. The system is also intensively applied for economic coding of speech sounds, which can be used for output with the commercially available speech synthesizer chip MEA8000.

Chapter 7, finally, outlines some possibilities for future research to further improve synthetic speech output of the system.

REFERENTIES

Atal B.S., David N. (1979) "On synthesizing natural sounding speech by linear prediction", Proc.IEEE ICASSP, 44-47.

Atal B.S., Hanauer S.L. (1971) "Speech analysis and synthesis by linear prediction of the speech wave", J.Acoust.Soc.Am. 50, 637-655.

Atal B.S., Remde J.R. (1982) "A new model of LPC excitation for producing natural sounding speech at low bit rates", Proc.IEEE ICASSP, 614-617.

Bierlaagh Th.C.J. (1982) MEA8000 application report, CAB report EDP 8201, Philips Elcoma, Eindhoven.

Bogaart P.C. Uit den (1975) Woordfrequenties in geschreven en gesproken Nederlands, Oosthoek Scheltema Hoekema, Utrecht.

Bouma H. (1976) Handboek der psychonomie, hfdst. 8, Van Loghum Slaterus, Deventer.

Brück H.E.van, Teuling D.J.A. (1982) "Integrated voice synthesizer", Electronic components and materials 4, 72-79.

Chiba T., Kajiyama M. (1958) The vowel, its nature and structure, Phonetic Society of Japan, Tokyo.

Darwin C.J. (1982) "Analysis and synthesis of mixed excitation LPC coded speech", IPO Ann.Progr.Rep. 17, 51-56.

Duifhuis H., Willems L.F., Sluyter R.J. (1982) "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception", J.Acoust.Soc.Am. 71, 1568-1580.

Egan J.P. (1948) "Articulation testing methods", Laryngoscope 58, 955-991. Herdrukt in: Hawley H.E.(ed) (1977) Speech intelligibility and speaker recognition, Dowden Hutchinson Ross, Stroudsburg, 175-206.

Eggermont J.P.M. (1956) "De klankfrequentie in het hedendaagse gesproken Nederlands", De nieuwe taalgids 49, 221-223.

Elsendoorn B.A.G., Hart J.'t (1982) "Exploring the possibilities of speech synthesis with Dutch phonemes", IPO Ann.Prog.Rep. 17, 63-65.

Essen H.A.van, Willems L.F. (1978) Formant coded voice response unit, Techn. Note 194/78, Philips Nat.Lab, Eindhoven.

Fairbanks G. (1958) "Test of phonemic differentiation: the rhyme test", J.Acoust.Soc.Am. 37, 158-166.

Fant G. (1960) Acoustic theory of speech production, Mouton, Den Haag.

Fant G. (1968) "Analysis and synthesis of speech processes" in: Malmberg B.(ed) Manual of phonetics, North Holland Publishing Comp., Amsterdam, 173-277.

Flanagan J.L. (1972) Speech analysis, synthesis and perception, Springer, Berlijn.

Fröberg C.E. (1969) Introduction to numerical analysis, Addison Wesley, Reading.

Goldstein J.L. (1973) "An optimal processor theory for the central

origin formation of pitch of compex tones", J.Acoust.Soc.Am. 54, 1496-1516.

Goldstein J.L., Gerson A., Srulovicz P., Furst M. (1978) "Verification of the optimal probabilistic basis of aural processing of pitch of complex tones", J.Acoust.Soc.Am. 63, 486-497.

Goodman D.J., Goodman J.S., Mun Chen (1973) "Consonant intelligibility and ratings of digitally coded speech", IEEE Trans.ASSP 26, 403-409.

Griffiths J.D. (1967) "Rhyming minimal contrasts: a simplified diagnostic articulation test", J.Acoust.Soc.Am 42, 235-248.

Grosjean F. (1980) "Spoken word recognition process and the gating paradigm", Perception & Psychophysics 28, 267-283.

Hart J.'t, Cohen A. (1973) "Intonation by rule: a perceptual quest", J. of Phonetics 1, 309-327.

Hart J.'t, Collier R. (1975) "Integrating different levels of intonation analysis", J. of Phonetics 3, 235-255.

Hart J.'t, Nooteboom S.G., Vogten L.L.M., Willems L.F. (1981/1982) "Manipulaties met spraakgeluid", Philips Techn. Tijdschr. 40, 108-119.

Holmes J.N. (1973) "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", IEEE Trans.AU 21, 298-305.

House A.S., Williams C.E., Hecker H.L., Kryter K.D. (1965) "Articulation testing methods: Consonantal differentiation with a closed response set", J.Acoust.Soc.Am. 37, 158-166.

IEEE (1969) "IEEE recommended practice for speech quality measurement."IEEE Trans.AU 17, 227-246.

Kalikow D.N., Stevens K.N., Elliott L.L. (1977) "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability", J.Acoust.Soc.Am. 61, 1337-1351.

Keeler L.O., Clement G.L., Strong W.J., Palmer E.P. (1976) "Two preliminary studies of intelligibility of predictor coefficient and formant coded speech", IEEE Trans.ASSP 24, 429-432.

Klatt D.H. (1976) "Structure of a phonological rule component for a synthesis-by-rule program", IEEE Trans.ASSP 24, 391-398.

Koopmans-van Beinum F.J. (1980) Vowel contrast reduction, Diss. Univ. Amsterdam.

Kroon J.N., Braal E.de (1980) "A talking typewriter as an aid for the visually impaired", IPO Ann.Progr.Rep.15, 118-123.

Makhoul J. (1975) "Linear prediction: a tutorial review", Proc.IEEE 63, 561-580.

Makhoul J., Viswanathan R., Schwartz R., Huggins A.F.W. (1978) "A mixed-source model for speech compression and synthesis", J.Acoust.-Soc.Am. 64, 1577-1581.

Markel J.D. (1972) "Digital inverse filtering, a new tool for formant

trajectory estimation", IEEE Trans.AU 20, 129-137.
Markel J.D., Gray A.H. (1976) Linear prediction of speech, Springer, Berlijn.
Marslen-Wilson W.D. (1980) "Speech understanding as a psychological process" in: Simon J.D. (ed) Spoken language generation and recognition, Reidel, Dordrecht, 39-68.
McGonegal C.M., Rabiner L.R., Rosenberg A.E. (1977) "A subjective evaluation of pitch detection methods using LPC synthesized speech", IEEE Trans.ASSP 25, 221-229.
Miller G.A., Nicely P.E. (1955) "An analysis of perceptual confusions among some English consonants", J.Acoust.Soc.Am. 27, 338-352.
Morton J. (1969) "Interaction of information in word recognition", Psychol.Review 76, 165-178.
Muller H.F. (1975) Procedures in P800 full Fortran, IPO rapport 276, IPO Eindhoven.
Nakatani L.H., Dukes K.D. (1973) "A sensitive test of speech communication quality", J.Acoust.Soc.Am. 53, 1083-1092.
Nooteboom S.G. (1972) Production and perception of vowel duration, Diss. Univ. Utrecht.
Nooteboom S.G. (1982) Fonetiek op het grensvlak tussen geluid en betekenis, Brill, Leiden.
Nooteboom S.G., Cohen A. (1976) Spreken en verstaan, Van Gorcum, Assen.
Nooteboom S.G., Doodeman G.D. (1982) "Speech quality and word recognition from fragments of spoken words", IPO Ann.Progr.Rep. 17, 46-50.
Nooteboom S.G., Doodeman G.D. (1983) "Speech quality and the gating paradigm" in: Cohen A., Broecke M.P.R.vd (eds) Abstracts of the tenth int. congress of phonetic sciences, Foris, Dordrecht, 535.
Nye P.W., Gaitenby J.H. (1973) "Consonant intelligibility in synthetic speech and in natural speech control (modified rhyme test results)", Haskins Laboratories, Status report on speech research SR-33, 77-91.
Olive J.P. (1977) "Rule synthesis of speech from dyadic units", Proc.-IEEE ICASSP, 568-570.
Olive J.P., Spickenagel N. (1976) "Speech resynthesis from phoneme related parameters", J.Acoust.Soc.Am. 59, 993-996.
Olive J.P., Liberman M. (1979) "A set of concatenative units for speech synthesis", in: Wolff J.J., Klatt D.H. (eds) ASA-50 Speech Comm. Papers, 515-518.
Plomp R. Mimpen A.M. (1979) "Improving the reliability of testing the speech reception threshold for sentences", Audiology 18, 43-52.
Pols L.C.W. (1977) Spectral analysis and identification of Dutch vowels in monosyllabic words, Diss. Univ. Amsterdam.
Pols L.C.W. (1979) Persoonlijke communicatie.

Polstra J. (1981) "Speech synthesizer interface", IPO Ann.Progr.Rep. 16, 143-144.

Pijper J.R.de (1983) Modelling British English intonation, Diss. Univ. Utrecht, Foris, Dordrecht.

Rabiner L.R., Schafer R.W. (1978) Digital processing of speech signals, Prentice Hall, London.

Rabiner L.R., Cheng M.J., Rosenberg J.A.E., McGonegal C.A. (1976) "A comparitive performance study of several pitch detection algorithms", IEEE Trans.ASSP 24, 399-418.

Slis I.H., Muller H.F. (1971) "A computer programme for synthesis by rule", IPO Ann.Progr.Rep. 6, 24-28.

Slis I.H., Nooteboom S.G., Willems L.F. (1977) "Speech synthesis by rule: An overview of a system used in IPO" in: Hambuger Phonetische Beiträge 22, Helmut Buske Verlag, Hamburg, 161-187.

Smith M.E., Robinson K.E., Strong W.J. (1981) "Intelligibility and qualitity of linear predictor and eigenparameter coded speech", IEEE Trans.ASSP 29, 391-395.

Steeneken H.J.M. (1982) Ontwikkeling en toetsing van een nederlandstalige rijmtest voor het testen van spraakkommunikatiekanalen, IZF rapport 1982-13, IZF, Soesterberg.

Vogten L.L.M. (1980) "Evaluation of LPC-formant coded speech with a speech interference test", IPO Ann.Progr.Rep. 15, 33-41.

Vogten L.L.M. (1983) In bewerking.

Vogten L.L.M., Willems L.F. (1977) "The Formator: a speech analysis and synthesis system based on formant extraction from linear prediction coefficients", IPO Ann.Progr.Rep. 12, 47-54.

Voiers W.D. (1977) "Diagnostic evaluation of speech intelligibility" in: Hawley M.E. (ed) Speech intelligibility and speaker recognition, Dowden Hutchinson Ross, Stroudsburg, 374-387.

Waterham R.P. (1983) Aanpassingen aan bloeddrukmeetapparatuur t.b.v. visueel gehandicapte diabetici, Vakgroep EME, afd. Elektrotechniek TH Eindhoven.

Willems L.F. (1976) "LPC analyse en formantsynthese van spraak" in: Transmissie en synthese van spraak, publ. nr.36, Ned. Akoestisch Genootschap, Delft, 27-34.

Willems L.F. (1982) Programs which implement the DWS pitch detector, IPO rapport 394, IPO, Eindhoven.

Willems L.F., Bierlaagh Th.C.J. (1983) "A formant synthesizer chip: the MEA8000" in: Cohen A., Broecke M.P.R.vd (eds) Abstracts of the tenth int. congress of phonetic sciences, Foris, Dordrecht, 396.

Willems L.F., Vogten L.L.M. (1979) Ned. Octrooi aanvrage nr. 7902631.

Willems N.J. (1982) English intonation from a Dutch point of view", Diss. Univ. Utrecht, Foris, Dordrecht.

Zelle M., Pijper J.R.de, Hart J.'t (1983) "Semi automatic synthesis of

intonation for Dutch and British English" in: Cohen A., Broecke M.P.R.vd (eds) Abstracts of the tenth int. congress of phonetic sciences, Foris, Dordrecht, 397.

CURRICULUM VITAE

15 sept. 1945	geboren te Kerkrade.
sept. 1957 - juli 1961	MULO A + B te Gulpen en Wijchen.
sept. 1961 - juli 1965	HTS elektrotechniek te 's-Hertogenbosch.
sept. 1965 - nov. 1971	TH elektrotechniek te Eindhoven. Afstudeeronderwerp: ultrasone flowmeting van gassen t.b.v. longfunktie-onderzoek.
sinds jan. 1972	Wetenschappelijk medewerker TH Eindhoven, gedetacheerd in het Instituut voor Perceptie-Onderzoek (IPO), voor het verrichten van onderzoek. Onderwerpen: tot medio 1977: auditieve maskering van zuivere tonen. sinds medio 1977: analyse, zuinige codering en resynthese van spraak.

STELLINGEN

1.
Bij gefluisterde spraak kunnen analyse en resynthese, zoals beschreven in de hoofstukken 2 en 3 van dit proefschrift, synthetische fluisterspraak opleveren die perceptief perfekt overeenkomt met het oorspronkelijke spraakgeluid.

2.
Voor synthese van perceptief bevredigende stemhebbende wrijfklanken in lopende spraak is gelijktijdige kombinatie van periodiek en ruisig brongeluid onnodig.

3.
De in Nooteboom en Cohen (1976) gewekte suggestie als zou voor het bemonsteren van signalen met uitsluitend positieve amplitudewaarden volstaan kunnen worden met slechts de halve Nyquistfrekwentie, is onjuist.

Nooteboom S.G., Cohen A. (1976) Spreken en verstaan, Van Gorcum, Assen, p.136.

4.
De gewoonte van sommige banken om na het nemen van hypotheken bij de renteberekening het jaar te stellen op 360 dagen maar de maand op het juiste aantal dagen, getuigt van bewuste misleiding van de cliënten.

5.
Het ontwikkelen en propageren van steeds weer - zogenaamd - nieuwe hogere programmeertalen, zonder aan te geven waarom of waarin de bestaande talen tekort zouden schieten, is verwerpelijk.

6.
Vooral binnen de bebouwde kom leveren sommige wegmarkeringen een negatieve bijdrage tot de verkeersveiligheid.

7.
Ook academische tradities kunnen verouderen.

Eindhoven, 18 november 1983 L.L.M. Vogten