Temporele decompositie van spraak, uitgaande van spectrale parameters

Instituut voor Perceptie Onderzoek
Postbus 513, 5600 MB Eindhoven

25.02.1988

Rapport no. 635

Temporele decompositie van spraak, uitgaande van spectrale parameters

L.W. Lemmens

Instituut voor Perceptie Onderzoek
Den Dolech 2
Eindhoven

Temporele decompositie van spraak, uitgaande van spectrale parameters.

L.W. LEMMENS

24 februari 1988

Verslag van een eerste stage.
Stagebegeleiding : Mw. Drs. A.M.L. van Dijk–Kappers en Ir. L.F. Willems.

Samenvatting

Atal heeft in 1983 een methode voorgesteld om een spraaksignaal op te delen in elkaar overlappende "gebeurtenissen", de zogenaamde "temporele decompositie". De "gebeurtenissen" zijn te relateren aan bewegingen van het spraakkanaal naar en van een bepaalde articulatorische doelpositie, en worden beschreven door doelfuncties en -vectoren.

In dit onderzoek is gekeken naar de fonetische relevantie van de resultaten van de decompositie, waarbij verschillende ingangsparameters zijn gebruikt. Een maat voor deze relevantie was het percentage gevallen waarin bij elke klank precies één doelfunctie gevonden werd. Bij het gebruik van spectrale amplitudecoëfficiënten als ingangsparameters bleek dat percentage rond de 70% te liggen. Formanten als ingangsparameters gaven een percentage van 73% te zien. Eerder onderzoek met log-area's had een resultaat van 67%, dat als goed beschouwd wordt.

Om perceptief na te gaan hoe goed de gevonden doelfuncties en -vectoren het oorspronkelijke signaal beschrijven is getracht om dat signaal te resynthetiseren, uitgaande van de doelfuncties en -vectoren. Dat bleek niet mogelijk vanwege het feit dat in het hele proces van analyse en decompositie de fase van het signaal steeds verwaarloosd werd.

Inhoudsopgave

Inhoud

Hoofdstuk 1

Inleiding

Het onderzoek naar menselijke en, daarmee samenhangend, computerspraak is de laatste jaren in opmars. In Nederland heeft het zogenaamde SPIN-project (Stimulerings Projectteam Informatica Nederland) dit jaar een budget van zes miljoen gulden gekregen voor de ontwikkeling van een compleet spraaksynthesesysteem, en in samenwerking met Philips en Siemens wordt aan het IPO bij de Akofon-groep onderzoek gedaan voor het zg. SPICOS project, een groot project voor het analyseren, interpreteren en produceren van spraak met behulp van computers.

In het kader van een onderzoek naar instrumenten voor spraakanalyse heb ik een stageonderzoek gedaan naar de bruikbaarheid van spectrale parameters als ingang voor temporele decompositie. Dit is een procedure die ontworpen is voor datareductie, maar erg bruikbaar is voor het onderzoeken van de temporele structuur van een spraaksignaal.

In hoofdstuk twee zal worden gekeken naar de manier waarop de parameters verkregen worden, de L.P.C. analyse. Dit is een methode om een spraaksignaal te analyseren, waarbij uitgegaan wordt van een bron-filter model van de menselijke spraakorganen. Deze methode levert parameters welke opgeslagen en later gebruikt kunnen worden voor de analyse en eventuele resynthese van het spraaksignaal. Er zal worden ingegaan op het bron-filtermodel en enige mathematische achtergronden van de L.P.C. analyse.

Omdat de opslag van de parameters veel computergeheugen kost heeft Atal [1] enkele jaren geleden een methode voor datareductie voorgesteld, die hij temporele decompositie heeft genoemd. Behalve datareductie blijkt deze methode ook een instrument te leveren om een de tijdstructuur van een spraakuiting te kunnen analyseren, en op die manier fonetisch relevante eenheden zoals fonemen te onderscheiden. De temporele decompositiemethode komt in hoofdstuk drie aan bod.

In het vierde hoofdstuk zal nader worden ingegaan op de onderzochte parameters, de opzet van het onderzoek en de resultaten ervan. Om perceptief na te gaan hoe goed de via de temporele decompositie verkregen doelfuncties en -vectoren het

oorspronkelijke signaal beschrijven is getracht het signaal te resynthetiseren. Ook hiervan wordt in hoofdstuk vier een beschrijving gegeven.

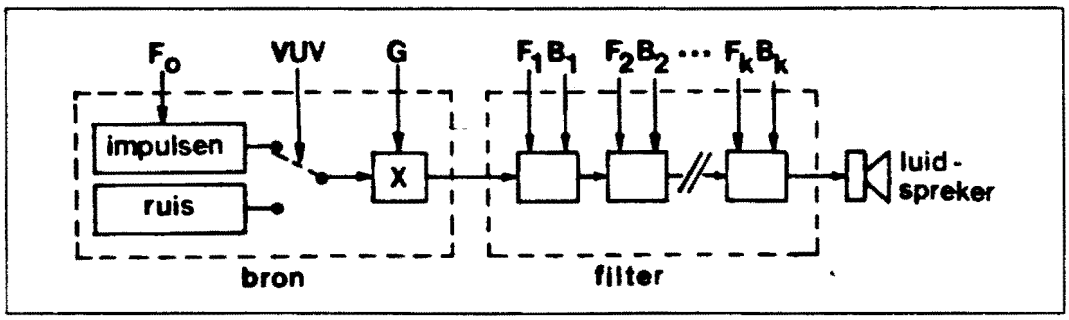
Hoofdstuk 2

Spraakanalyse en -synthese

2.1 Inleiding

Spraakanalyse en -synthese zijn gebieden die zich in een groeiende belangstelling mogen verheugen. Omdat spraak voor mensen een communicatiemiddel is dat uitermate geschikt is waar snelheid van belang is, of waar de handen niet vrij zijn, zou computerspraak een belangrijk hulpmiddel voor de toekomst kunnen zijn.

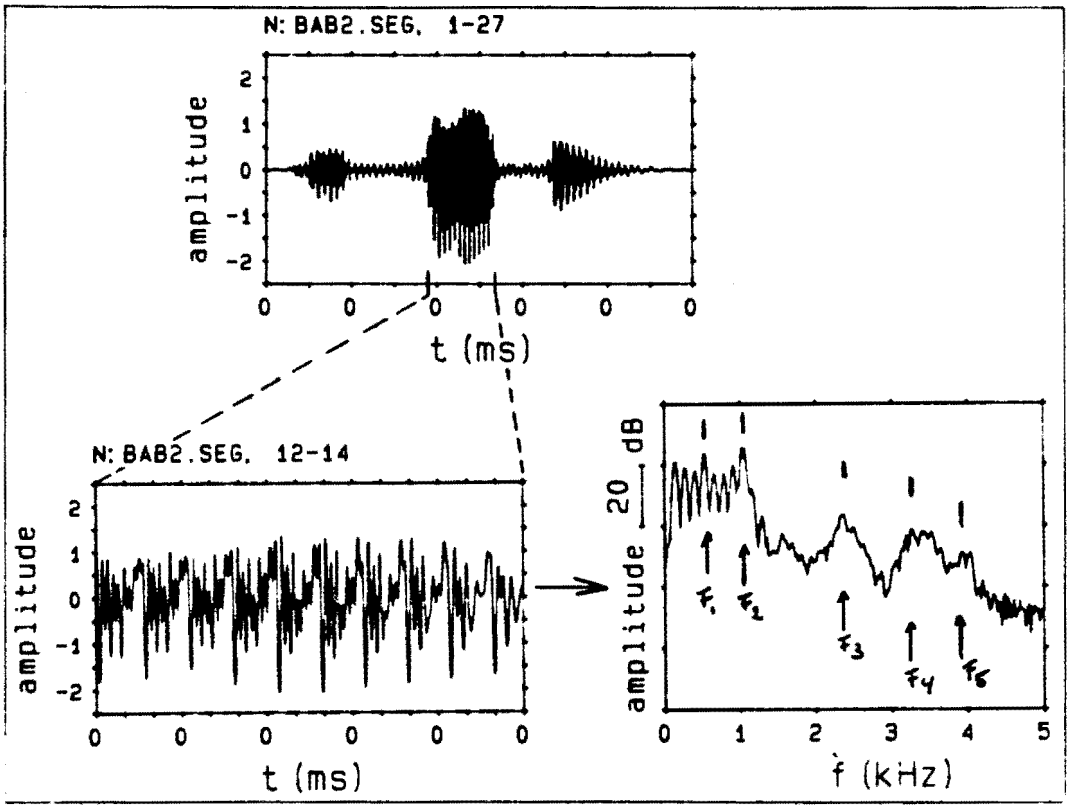
Verder is spraakanalyse een belangrijk hulpmiddel voor de fonetiek, daar men op die manier kan onderzoeken hoe menselijke spraak werkt. Een analyse-resynthesemethode is de zogenaamde *L.P.C. analyse*, waarbij uitgegaan wordt van een *bron-filter model* (figuur 2.1).



Figuur 2.1: Vereenvoudigd bron-filter model voor de productie van spraakgeluid. Het bronsignaal wordt bepaald door drie parameters : de herhalingsfrequentie f_0 van de impulsen, de stem/stemloosparameter V/UV en de versterkingsfactor G. Het filter is samengesteld uit een cascade van 2^e-orde deelfilters. De parameters van ieder filter zijn aangeduid met afstemfrequenties F_k en bandbreedtes B_k.

2.2 Het bron-filter model

Voor de beschrijving van spraakgeluid onderscheiden we een *geluidsbron*, en een *filter* waardoor het brongeluid wordt gekleurd. Dit model heet het bron-filter model. In de normale spraak zijn er twee typen brongeluiden : *stembandtrillingen* voor productie van stemhebbende klanken, en *ruisgeluid* voor stemloze spraakklanken. Het filter wordt gevormd door de mond-keelholte. Hier wordt het bronsignaal zó vervormd, dat er in het spectrum van de spraakklank gebieden van relatief hoge energiedichtheid te onderscheiden zijn (zie figuur 2.2). Deze relatieve maxima (zg. *formanten*) zijn met name karakteristiek voor afzonderlijke klinkers en tweeklanken. Ook in de analyse



Figuur 2.2: Voorbeeld van de golfvorm van het woordje debabe *(/dəbabə/). Eronder staan de golfvorm van de klinker /a/, en het spectrum ervan. In het spectrum staan de formanten aangegeven (F1 - F5).*

en resynthese van spraak maakt men gebruik van het bron-filter model (zie figuur 2.1). Het model bestaat hier uit een geluidsbron waarmee zowel een harmonisch signaal als ruis opgewekt kan worden, en een hogere orde filter dat de kleuring van het geluid voor zijn rekening neemt.

2.3 De bepaling van het filter met behulp van lineaire predictie

De parameters voor het filter worden bepaald via een methode die lineaire predictie genoemd wordt. In het algemeen kan een (discreet) signaal s_n beschouwd worden als de uitgang van een systeem met een onbekend (discreet) ingangssignaal u_n, zó dat de volgende relatie geldt :

$$s_n = -\sum_{k=1}^{p} a_k s_{n-k} + \mathcal{G} \sum_{l=0}^{q} b_l u_{n-l}, \qquad b_0 = 1 \tag{2.1}$$

waarbij a_k, b_l, en $\mathcal{G}$ de parameters van het hypothetische systeem zijn. Volgens vergelijking 2.1 zijn de uitgangssignalen s_n een lineaire functie van de ingangssignalen u_n en de voorgaande uitgangssignalen. Dus s_n is ***voorspelbaar*** uit lineaire combinaties van voorafgaande in- en uitgangssignalen. Vandaar de naam ***lineaire predictie*** (LPC betekent Linear Predictive Coding).

In het frequentiedomein wordt de overdracht van het filter gegeven door :

$$H(z) = \frac{\mathcal{G}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.2}$$

Hierbij is $\mathcal{G}$ een versterkingsfactor. Het filter wordt door $H(z)$ volledig bepaald. Deze manier van beschrijven heet het *all-pole* model, vanwege het feit dat de overdracht door z'n polen beschreven wordt.

2.4 Schatting van de parameters van het filter

We veronderstellen nu dat het ingangssignaal totaal onbekend is. We kunnen het uitgangssignaal s_n nu alleen bij benadering voorspellen uit een lineair gewogen som van voorgaande bemonsteringen. Noemen we deze benadering van het signaal $\tilde{s}_n$, dan gaat vergelijking 2.1 over in :

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k s_{n-k}. \tag{2.3}$$

De fout tussen de geschatte waarde $\tilde{s}_n$ en de werkelijke waarde s_n wordt dan gegeven door :

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k s_{n-k}. \tag{2.4}$$

Het getal e_n wordt het foutsignaal of ook het residu genoemd. We bepalen nu de parameters van het filter via de kleinste kwadraten methode, waarbij we de totale kwadratische fout $E = \sum e_n^2$ minimaliseren op het oneindig lange interval $-\infty < n < \infty$. Daartoe stellen we :

$$\frac{\partial E}{\partial a_i} = 0, \qquad 1 \leq i \leq p. \tag{2.5}$$

Het voorgaande stelsel gaat dan vervolgens over in :

$$\sum_{k=1}^{p} a_k R(i-k) = -R(i), \qquad 1 \leq i \leq p \tag{2.6}$$

waarbij :

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \tag{2.7}$$

de autocorrelatiefunctie van het uitgangssignaal s_n is. De minimale, totale kwadratische fout E_p wordt nu :

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{2.8}$$

Ook als s_n slechts over een beperkte tijdsduur bekend is, is deze methode toe te passen. Daarvoor moet s_n wel eerst met een *vensterfunctie* vermenigvuldigd worden, zodat de waarden voor $n < 0$ en $n > N$ nul zijn. De autocorrelatiefunctie 2.7 gaat dan over in :

$$R(i) = \sum_{n=0}^{N-i} s_n s_{n+i} \tag{2.9}$$

Er zijn verschillende zuinige en stabiele algorithmen om uit de voorgaande vergelijkingen de parameters a_k te bepalen. Verder is aan te tonen dat de versterkingsfactor $\mathcal{G}$ gelijk is aan :

$$\mathcal{G}^2 = E_p = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{2.10}$$

De hierboven beschreven methode van LPC-analyse heet de autocorrelatiemethode en geeft (in tegenstelling tot andere methodes) altijd een stabiel filter : alle polen van H(z) in vergelijking 2.2 liggen binnen de eenheidscirkel in het complexe vlak, en de autocorrelatiecoëfficiënten R(i) in vergelijking 2.9 zijn definiet positief. [2,5,8,11].

2.5 De LPC-analyse in de praktijk

Met behulp van de LPC-analyse kunnen nu de polen van de overdrachtsfunctie 2.2 berekend worden. De in het onderzoek toegepaste programmatuur verdeelde het signaal in frames met een duur van 25 ms, die telkens 10 ms verschoven zijn. De bemonsteringsfrequentie was 10000 Hz, er werden per frame 10 of 16 a_k parameters

bepaald, samen met de versterkingsfactor $\mathcal{G}$, de V/UV - parameter, en de bronfrequentie F_0. Al deze gegevens werden opgeslagen in zogenaamde a/p-files.

Het is mogelijk om de filterparameters om te rekenen in andere parameters die het filter ook beschrijven. Viswanathan en Makhoul [13] hebben een aantal parametersets die hiervoor in aanmerking komen beschreven. De gebruikte programmatuur slaat de parameters op in de a/p-files, en doet dit naar wens in de vorm van a_k-parameters (de polen van één k^e orde filter), p/q-parameters (de polen van een cascade van $\frac{1}{2}k$ tweede-orde filters), r-parameters (reflectiecoëfficiënten) of F/B-parameters (formanten en bandbreedtes) in zogenaamde a/p-files. Uitgaande van deze files kunnen dan verdere bewerkingen uitgevoerd worden.

Hoofdstuk 3

Temporele decompositie

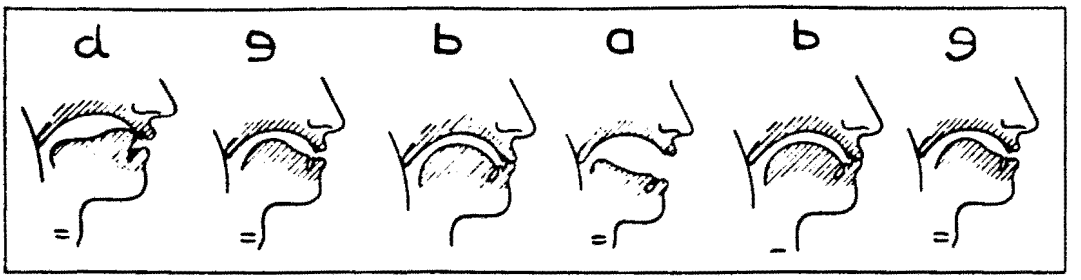
3.1 Inleiding

De spraakparameters zoals ze door de LPC-analyse geleverd worden bevatten nog een grote hoeveelheid redundante informatie. Een gevolg hiervan is dat spraakproductie met behulp van deze parameters een hoge communicatiesnelheid van een computer vereisen, samen met een hoge opslagcapaciteit. *Temporele decompositie* biedt een mogelijkheid om zuiniger te coderen. Omdat de temporele decompositie informatie geeft over de temporele opbouw van een spraaksignaal kan de methode ook gebruikt worden voor de analyse van fonetisch relevante elementen in het signaal.

3.2 Theorie

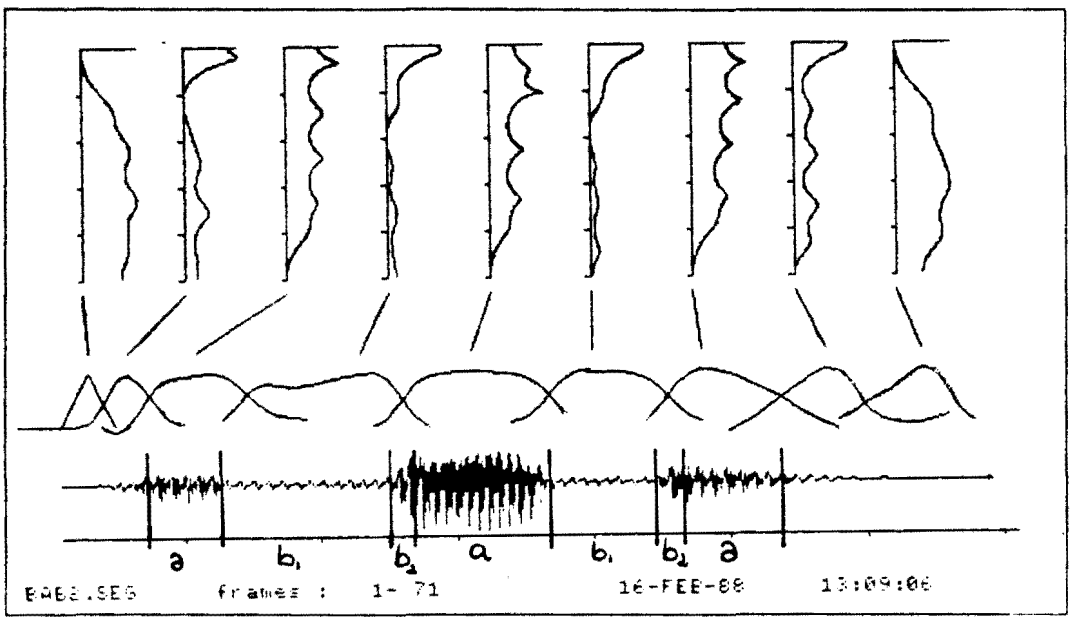
De productie van spraak gebeurt in tijdsintervallen van variabele lengte. Er zijn articulatorische bewegingen die vrij traag zijn, maar er zijn er ook die relatief snel zijn. Uniforme bemonstering van spraak is dus niet erg efficient. Atal [1] heeft in 1983 een methode voor economischer codering van LPC-parameters voorgesteld, de zg. *temporele decompositie*. Dit is een procedure om na de LPC-analyse, uit de continue verandering van spraakparameters discrete eenheden van variabele lengte te berekenen. De opslag van deze eenheden biedt een zuiniger mogelijkheid om spraak te coderen dan de opslag van LPC-parameters. Verder geven deze eenheden, de zg. articulatorische doelfuncties, of weegfuncties, een indicatie van de temporele opbouw van een spraaksignaal.

Articulatorische fonetiek beschouwt spraak als een opeenvolging van overlappende "bewegingen". De overlap zorgt voor de karakteristieke overgang tussen fonemen. Dit resulteert in de bewegingen van het mond-keelholte kanaal (tong, lippen ed.) van de ene articulatorische positie naar de andere, tijdens het uitspreken van een woord (zie figuur 3.1). In de temporele decompositie wordt elke articulatorische doelpositie beschreven door een *doelvector*, $a_i(k)$, en de beweging zelf door



***Figuur 3.1:** Schematische weergave van de stand van het mond-keelholtekanaal tijdens de spraakuiting* debabə *(/dəbabə/).*

een tijdsafhankelijke *doelfunctie* $\phi_k(n)$ (zie figuur 3.2). Elke functie $\phi_k(n)$ is slechts in een relatief kort tijdsinterval ongelijk aan nul. De aanname van Atal is nu dat



***Figuur 3.2:** Voorbeeld van de indeling van de spraakuiting* debabə *(/dəbabə/) in doelfuncties $\phi_k(t)$ en de bijhorende akoestische doelvectoren $\vec{a}_k$.*

we de met behulp van de L.P.C. analyse verkregen spraakparameters $y_i(n)$ kunnen benaderen door een lineaire combinatie van doelvectoren en -functies :

$$y_i(n) = \sum_{k=1}^{m} a_i(k)\phi_k(n) \qquad 1 \leq i \leq p. \tag{3.1}$$

Hierbij is m het totale aantal articulatorische bewegingen in de spraakuiting, en p het aantal parameters dat bij de bemonstering bepaald wordt. Als we vergelijking 3.1 inverteren, dan krijgen we elke ϕ als lineaire combinatie van de ingangsparameters

y. Bepalen we van de matrix y de principale componenten u met behulp van een zg. ***singuliere waarden decompositie***, dan krijgen we een datareductie door de λ meest significante waarden te gebruiken in plaats van y :

$$\phi(n) = \sum_{i=1}^{\lambda} b_i u_i(n). \tag{3.2}$$

Omdat iedere ϕ-functie een beweging naar, en daarna van, een bepaalde doelpositie weergeeft, moet hij "compact" zijn in de tijd. Dit wil zeggen dat elke ϕ-functie lange tijd nul is, nabij het doel geleidelijk één wordt, en vervolgens weer afneemt tot nul. De waardes u_i in vergelijking 3.2 zijn bekend, en de waardes b_i moeten nu zó berekend worden, dat $\phi(n)$ aan bovenstaande voorwaarden voldoet. Verschillende methodes hiervoor zijn gepubliceerd door Atal [1], en van Dijk-Kappers en Marcus [4]. Marcus en van Lieshout [10] bespreken de nadelen van de methode van Atal.

3.3 De keuze van parameters

Omdat de temporele decompositie iets zegt over de tijdstructuur van een spraaksignaal is het door keuze van een goede parameterset mogelijk om de methode te gebruiken voor de analyse van spraak in fonetisch relevante eenheden. Voor een goede parameterset geldt dan dat de door temporele decompositie gevonden weegfuncties een fonetische relevantie hebben. Viswanathan en Makhoul [13] hebben een aantal parametersets voor de LPC-analyse voorgesteld. Atal [1] heeft de temporele decompositie voorgesteld als methode voor datareductie, gebruik makend van log-area's. Benning [3] heeft log-area's, area's, en reflectiecoëfficiënten op hun bruikbaarheid voor temporele decompositie onderzocht.

In dit onderzoek is gekeken naar de bruikbaarheid van een tweetal spectrale parameters. Als maat voor de bruikbaarheid geldt het percentage van de gevallen waarin een foneem door precies één weegfunctie weergegeven wordt. Voor de log-area parameters was dat percentage rond de 67%, dat als goed beschouwd wordt. De spectrale parameters in dit onderzoek, zijn spectrale amplitudecoëfficiënten, en formanten. Daar geen van beide parameters snelle variaties in de tijd vertonen, werd verwacht dat ze goed bruikbaar zouden zijn als hulpmiddel voor het detekteren van fonemen in een spraaksignaal.

Hoofdstuk 4

Het onderzoek

4.1 Inleiding

Het bron-filter model beschouwt het mondkanaal als een filter dat het bronsignaal van de stembanden kleurt. Dat wil zeggen dat bij elke articulatorische positie van het mondkanaal een specifiek spectrum hoort. Nadat via een LPC-analyse de overdrachtskarakteristiek van het filter bepaald is, kunnen via een Fouriertransformatie de amplitudecoëfficiënten van dit spectrum berekend worden. Ook kunnen na de LPC-analyse, uitgaande van autocorrelatiecoëfficiënten, formanten en bandbreedtes bepaald worden die het filter beschrijven. De zo verkregen amplitudecoëfficiënten en formanten kunnen als ingang van een temporele decompositie gebruikt worden.

4.2 Opzet van het onderzoek

4.2.1 De onderzochte parameters

De parameters in dit onderzoek zijn spectrale parameters. Als eerste zijn de amplitudecoëfficiënten onderzocht. De LPC-analyse verdeelt het signaal in frames met een duur van 25ms die telkens 10ms verschoven worden. Voor elk frame worden de a_k-waarden uit vergelijking 2.2 bepaald. Deze waarden beschrijven de temporele overdrachtskarakteristiek van het filter (uit het bron-filter model). Een discrete Fouriertransformatie van deze waarden geeft k amplitudecoëfficiënten die de spectrale overdrachtskarakteristiek van het filter beschrijven. Deze komt dan overeen met het spectrum van het signaal. In het onderzoek zijn de waarden $k = 10$ en $k = 16$ gebruikt.

In het tweede onderzoek zijn formanten gebruikt als parameters. Hier is de formantanalyse methode van Willems [14] toegepast. Deze methode gebruikt een zogenaamde split Levinson recursie om een set van vijf formantfrequenties met bijhorende bandbreedtes te berekenen, en heeft als voordeel boven andere methodes dat onder alle omstandigheden vijf waarden gevonden worden. Het programma dat

op deze methode gebaseerd is schrijft de formant frequenties in een data-file zodat ze direct in te lezen zijn voor de temporele decompositie.

4.2.2 De analyse

In navolging van eerder onderzoek van onder andere Benning [3] waarbij area-, log-area-, en reflectiecoëfficiënten werden bekeken, is de temporele decompositie uitgevoerd op een serie van 47 zogenaamde CVC-woorden (Consonant, Vokaal, Consonant). De CVC-woorden bestaan uit een korte klinker *a*, *i* of *o*, omsloten door de medeklinkers *b*, *p*, *l* of *m*. Ten behoeve van een stabiele klankomgeving worden de zo verkregen lettergroepen voorafgegaan door een /də/-klank (klinkt als "de" in "de boom") en gevolgd door een /ə/ (dit is een zogenaamde schwa en klinkt als de "e" in het woordje "de"). Zo ontstaat door variatie van de begin- en eindmedeklinker en de klinker van de middengroep een serie betekenisloze woorden zoals *debabe* (/dəbabə/). Voor de plofklanken /b/ en /p/ geldt dat ze op te delen zijn in twee stukken : een korte stilte en de plof. Deze stukken noemen we respectievelijk *b1 en b2*, en *p1 en p2*.

Met behulp van een LPC-analyse zijn de filter-parameters (de a_k-parameters in vergelijking 2.2) van deze woorden bepaald en opgeslagen in databestanden, waarbij voor de analyse een bemonsteringsfrequentie van 10000 Hz gold, en er per frame 10 of 16 parameters bepaald werden (zie ook hoofdstuk 2). Bestaande programmatuur voor het uitvoeren van de decompositie bewerkte de LPC-parameters zo dat log-areacoëfficiënten verkregen werden. De programma's zijn zó herschreven dat de spectrale parameters voor temporele decompositie beschikbaar kwamen.

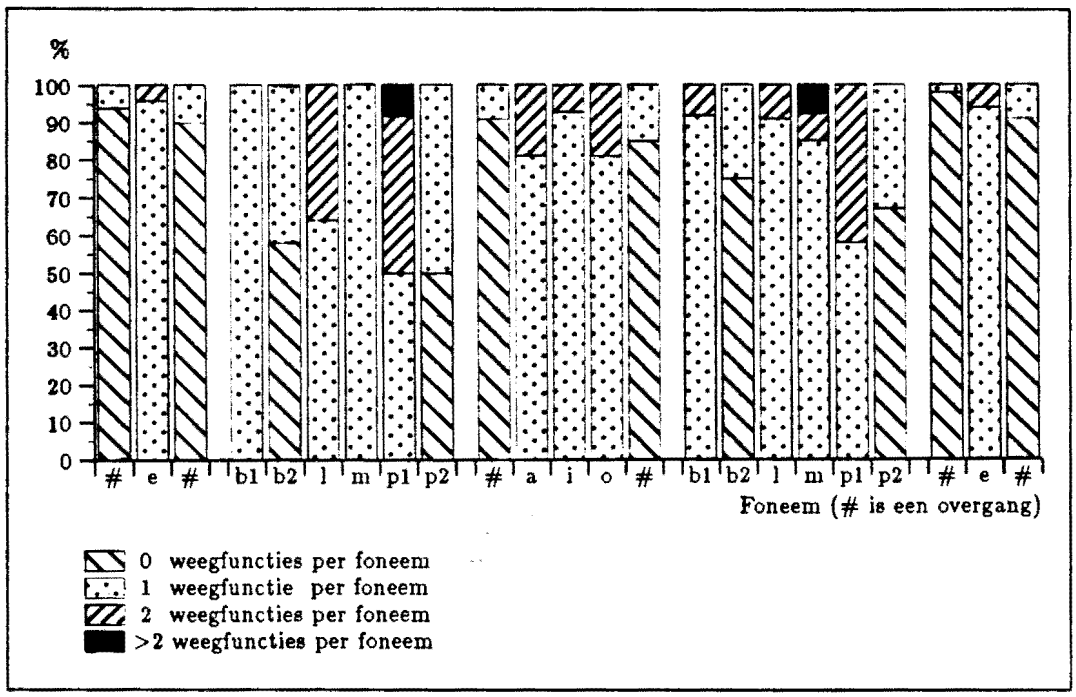
Voor het onderzoek van de parameters is de decompositie met behulp van het programma uitgevoerd, en de zo verkregen weegfuncties zijn opgeslagen in een databestand. Met behulp van een volgend programma, dat ook is aangepast, zijn de bijhorende akoestische doelvectoren berekend en aan het databestand toegevoegd. De verdere uitvoer van dit programma bestond uit een figuur waarin de golfvorm van het woord, de weegfuncties en de bijhorende akoestische doelvectoren werden weergegeven (zie figuur 3.2).

Nadat de hele serie CVC-woorden met bovengenoemde programma's was bewerkt, zijn per woord de grenzen tussen de fonemen op het gehoor bepaald en in het figuur aangegeven. Daarna is geteld hoeveel weegfuncties er per foneem gevonden zijn. Zo kon er een overzicht gemaakt worden van het aantal weegfuncties dat het programma per foneem vond. Als maat voor de bruikbaarheid van een parameterset is het percentage genomen dat weergeeft hoe vaak een willekeurig foneem door precies één weegfunctie beschreven wordt. Er waren 47 woorden, met in totaal 192 fonemen (daarbij zijn alleen de *b1, b2, l, m, p1, p2*, de *a, i, o*, en de *b1, b2, l, m, p1, p2* gerekend). De /ə/-fonemen zijn niet meegenomen in de bepaling van bovengenoemde maat, maar zijn wel geteld omdat ze bijdragen tot het algemene beeld. Van elke meting is een histogram gemaakt waarin per foneem is aangegeven hoe vaak er

een bepaald aantal weegfuncties is gevonden.

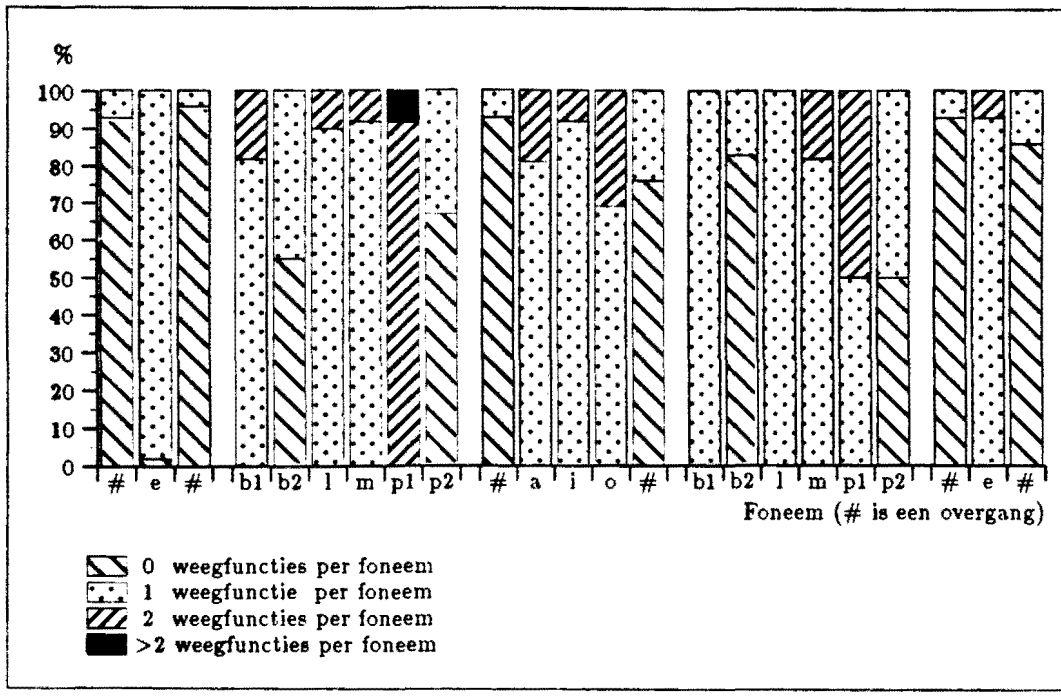
4.3 De resultaten van het onderzoek

De resultaten van het onderzoek zijn hieronder weergegeven in de vorm van histogrammen. Voor de interpretatie van de histogrammen moet in het oog gehouden worden hoe de waarden zijn weergegeven. Elke balk geeft exact 100% weer, en is ingedeeld naar het aantal weegfuncties dat per foneem gevonden wordt. In figuur 4.1 bijvoorbeeld wordt voor de *p1* in 50% van de gevallen precies één weegfunctie gevonden, en worden in 41% van de gevallen 2 weegfuncties gevonden, en in 9% van de gevallen meer dan twee.

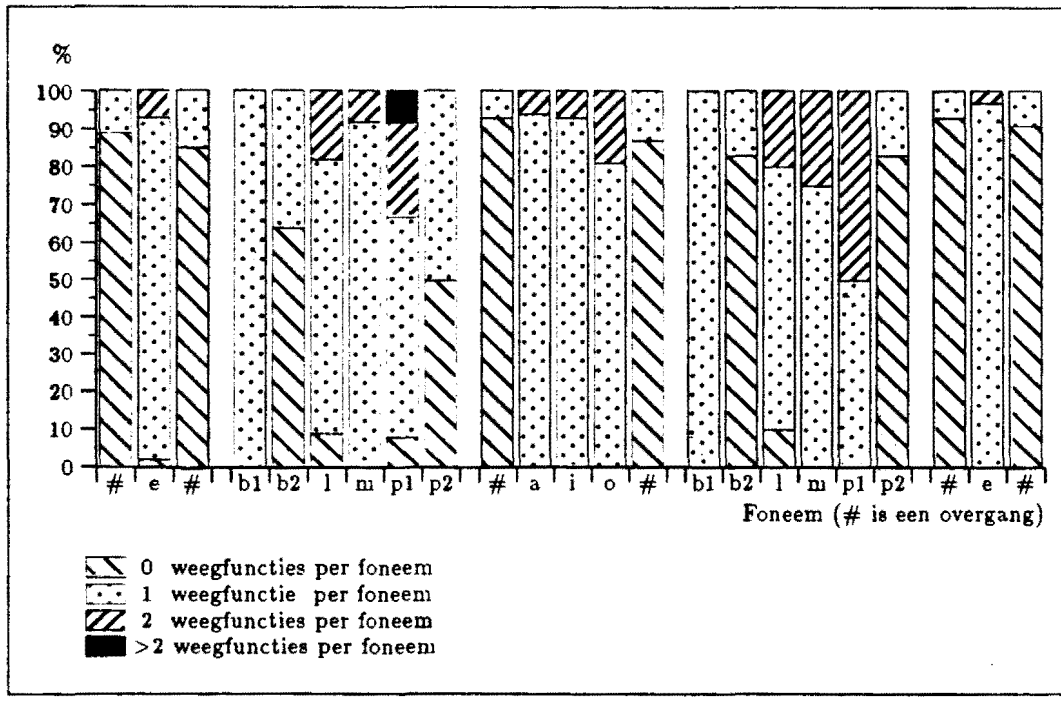


Figuur 4.1: Procentueel aantal gevallen waarin een foneem door respectievelijk 0, 1, 2, of meer dan 2 weegfuncties beschreven wordt, weergegeven voor de verschillende fonemen, uitgaande van 10 amplitudecoëfficiënten per frame en een vierde orde Fouriertransformatie.(# is een overgang tussen twee opeenvolgende fonemen of tussen een stilte en een foneem)

De figuren 4.1, 4.2 en 4.3 geven de resultaten van drie verschillende metingen met dezelfde soort coëfficiënten. Bij de tweede meting is alleen het aantal coëfficiënten per frame verhoogd, en bij de derde is de orde van de Fouriertransformatie één hoger gemaakt. Te zien is dat deze veranderingen niet veel invloed hebben op het uiteindelijke resultaat. De verwachting was dat het gebruik van meer parameters

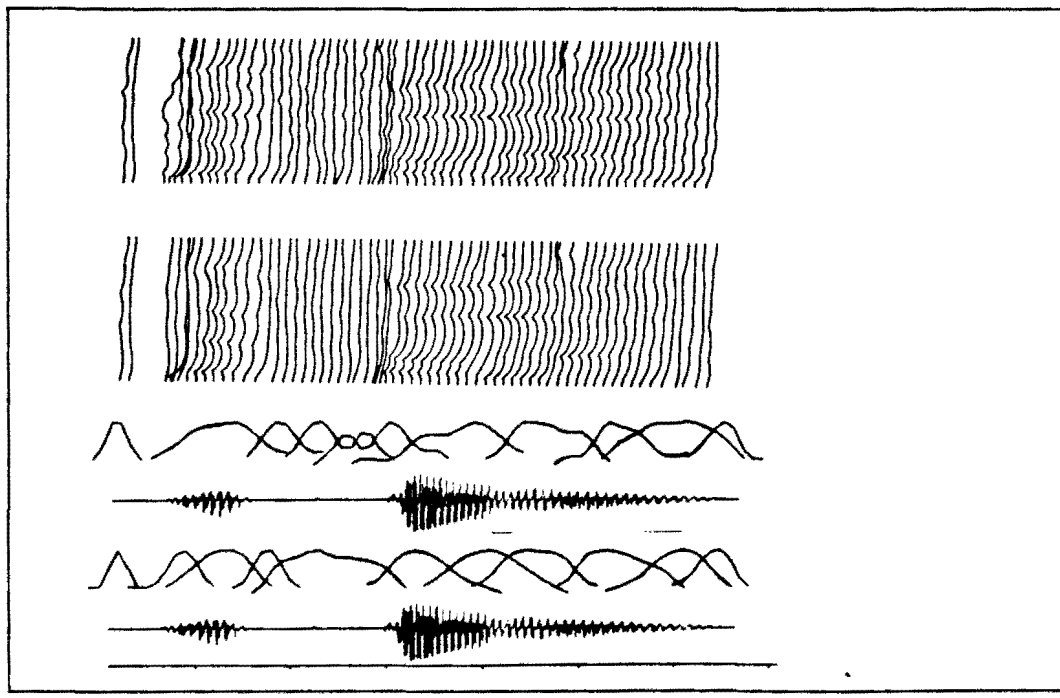


Figuur 4.2: Procentueel aantal gevallen waarin een foneem door respectievelijk 0, 1, 2, of meer dan 2 weegfuncties beschreven wordt, weergegeven voor de verschillende fonemen, uitgaande van 16 amplitudecoëfficiënten per frame en een vijfde orde Fouriertransformatie.



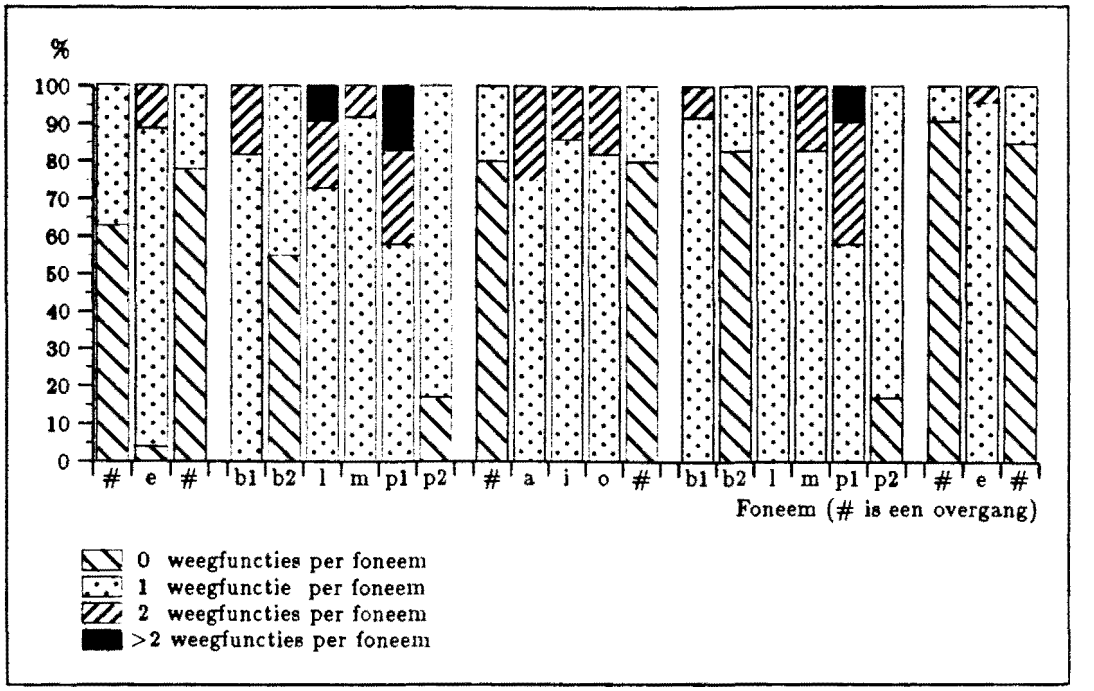
Figuur 4.3: Procentueel aantal gevallen waarin een foneem door respectievelijk 0, 1, 2, of meer dan 2 weegfuncties beschreven wordt, weergegeven voor de verschillende fonemen, uitgaande van 16 amplitudecoëfficiënten per frame en een zesde orde Fouriertransformatie.

per frame zou leiden tot een verbetering van de beschrijving door weegfuncties, en dat het verhogen van de orde van de DFT nog een verbetering zou opleveren. Een vergelijking van het eerste histogram met de andere geeft te zien dat er inderdaad een geringe verbetering is voor wat betreft het ontdekken van het b2-foneem, het p2-foneem, en het l-foneem. Gebruik makend van eerdergenoemde maat blijkt echter dat de situatie als geheel verslechtert. In het eerste geval wordt 71% van de fonemen door één weegfunctie weergegeven, in het tweede geval 69% en in het derde geval 67%.
Bij het zoeken naar een verklaring voor deze verslechtering moet bedacht worden dat de temporele decompositie erg gevoelig is voor variaties van het ingangssignaal. Dit ingangssignaal staat voor het woordje *depome* (/dəpomə/) weergegeven in figuur 4.4, samen met de golfvorm en weegfuncties. Te zien is dat het ingangssignaal veel



Figuur 4.4: Golfvorm, ingangssignaal en weegfuncties behorende bij het woordje /dəpomə/. Horizontaal loopt de tijdas en de verticale as is de frequentieas.

meer variatie vertoont als er 16 parameters per frame gebruikt worden. Daardoor wordt het signaal dan wel nauwkeuriger beschreven, maar het heeft ook tot gevolg dat er meer weegfuncties per foneem gevonden worden, en het percentage gevallen waarin precies één weegfunctie gevonden wordt, afneemt. Maar als geheel genomen is het resultaat goed te noemen en mag de conclusie getrokken worden dat de spectrale amplitudecoëfficiënten bruikbaar zijn voor het verdelen van een stuk spraak in fonemen.



Figuur 4.5: Procentueel aantal gevallen waarin een foneem door respectievelijk 0, 1, 2, of meer dan 2 weegfuncties beschreven wordt, weergegeven voor de verschillende fonemen, uitgaande van 5 formanten per frame.

Figuur 4.5 geeft de resultaten van de temporele decompositie met formanten als parameters. Het totale resultaat was hier zelfs nog iets beter : 73% van de fonemen werd door precies één weegfunctie beschreven. Vooral bij de p2-fonemen van de korte plofklank */p/* is een grote verbetering te zien in vergelijking met de amplitudecoëfficiënten. Ook deze parameterset mag als analyse instrument zeer bruikbaar genoemd worden.

4.4 Resynthese van het signaal

Om perceptief na te gaan hoe goed de weegfuncties en doelvectoren het oorspronkelijke signaal beschrijven, en om eventuele verdere statistische bewerkingen uit te kunnen voeren, is getracht de woorden, uitgaande van de weegfuncties en doelvectoren die met de eerste parameterset gevonden zijn, te resynthetiseren en hoorbaar te maken. Omdat daarvoor geen programmatuur voorhanden is, maar wel voor het het hoorbaar maken van a/p-files (dit zijn de databestanden die de bron en filterparameters uit de LPC-analyse bevatten), is besloten om de volgende strategie te gebruiken :

1. De berekening van een spectrum uit de weegfuncties en bijbehorende doelvectoren met behulp van de methode van Atal $\vec{y} = \sum_{k=1}^{m} \vec{a_k}\phi_k$ (zie Hoofdstuk 3).
2. Vervolgens de zo verkregen amplitudecoëfficiënten kwadrateren om een powerspectrum te krijgen.
3. De berekening van de autocorrelatiecoëfficiënten van het signaal uit het powerspectrum via een inverse Fourier transformatie.
4. Het bepalen van filterparameters (a_k in vergelijking 2.2) uit autocorrelatiecoëfficiënten door middel van een zogenaamde Levinson recursie.
5. Ten slotte de parameters in het juiste format opslaan in een a/p-file, en de woorden hoorbaar maken met behulp van de bestaande programmatuur.

Er is een routine geschreven om de achtereenvolgende stappen uit te voeren (zie appendix A). De procedures voor punt 4 en 5 bestonden al, de andere twee zijn geschreven en uitvoerig getest. Daarbij kwam al direkt een moeilijkheid naar voren: bij het bepalen van de amplitudecoëfficiënten door middel van een Fouriertransformatie wordt verder geen rekening meer gehouden met de fase van het signaal. Later bij het omzetten van het powerspectrum in autocorrelatiecoëfficiënten in stap 3 wordt de door de inverse Fouriertransformatie berekende fase weer verwaarloosd. Uit een test is gebleken dat daardoor het signaal flink vervormd werd (een beschrijving van deze test vindt u in de appendix). Het gevolg hiervan is dat de autocorrelatiecoëfficiënten niet allemaal positief definiet meer zijn, wat een vereiste is voor de stabiliteit van het filter (zie vergelijkingen 2.6 - 2.7). De Levinson recursie uit stap 4 bepaalt nu filtercoëfficiënten van een instabiel filter. De procedures die bij de resynthese gebruikt worden kunnen alleen stabiele filters bewerken, zodat de resynthese mislukte.

Een resynthese van het signaal uitgaande van de via temporele decompositie van formanten verkregen data is van de hand gewezen omdat de formanten het filter maar gedeeltelijk beschrijven (de bij de formanten horende bandbreedtes zijn niet in de temporele decompositie meegenomen).

4.5 Conclusie

Uit het onderzoek blijkt dat temporele decompositie uitgaande van spectrale parameters goed bruikbare resultaten oplevert wat betreft het opdelen van spraak in fonemen. De resultaten bij het gebruik van formanten iets beter dan bij het gebruik van spectrale amplitudecoëfficiënten. Naarmate de amplitudecoëfficiënten meer redundante informatie gaan bevatten worden ze minder bruikbaar voor het gestelde doel.

De resynthese van het signaal uitgaande van spectrale amplitudecoëfficiënten blijkt moeilijkheden te geven en dient nader onderzocht te worden.

Referenties

[1] B.S. Atal.
Efficient coding of LPC parameters by temporal decomposition. (1983),
Proceedings ICASSP-83, 2.6, 81-84.

[2] B.S. Atal and S.L. Hanauer.
Speech analysis and synthesis by linear prediction of the speech wave. (1971),
The Journal of the Acoustical Society of America, vol. 50, no.2, part 2, 637-655.

[3] F.J. Benning.
Temporele decompositie van spraak, uitgaande van andere akoestische parameters dan log-areas. (1987),
IPO rapport no. 593.

[4] A.M.L. van Dijk-Kappers and S.M. Marcus.
Temporal decomposition of speech. (1987),
IPO annual progress report 22, 1987.

[5] J.L. Flanagan.
Speech analysis, synthesis and perception. (1965),
Kommunikation und Kybernetik in Einzeldarstellungen, Band 3.

[6] J. 't Hart, ea.
Compendium college "Spraaktechnologie". (1987).

[7] J. 't Hart, S.G. Nooteboom, L.L.M. Vogten en L.F. Willems.
Manipulaties met spraakgeluid. (1981-82),
Philips Technical Reviews 40, 108-119.

[8] J. Makhoul.
Linear prediction : A tutorial review. (1975),
Proceedings of the IEEE, vol. 63, no. 4, 561-580.

[9] J. Makhoul.
Spectral analysis of speech by linear prediction. (1973),
IEEE, Transaction on audio electro-acoustics, vol.AU-21, 140-148.

[10] S.M. Marcus and R.A.J.M. van Lieshout.
Temporal decomposition of speech. (1984),
IPO Annual Progress Report 19, 25-31, 1984.

[11] C.D. McGillem and G.R. Cooper.
Continuous and discrete signal and system analysis. (1974),
Holt, Rinehart and Winston series in Electrical Engeneering, Electronics, and Systems.

[12] S.G. Nooteboom and A. Cohen.
Spreken en verstaan. Een nieuwe inleiding tot de experimentele fonetiek. (1984).

[13] R. Viswanathan and J. Makhoul.
Quantization properties of transmission parameters in linear predictive systems. (1975),
IEEE Transactions on acoustics, speech, and signal processing, vol. ASSP-23, no. 3, 309-321.

[14] L.F. Willems.
Robust formant analysis. (1986),
IPO annual progress report 21, 1986.

Appendix A

De resynthese van het signaal

Zoals in hoofdstuk vier beschreven heb ik getracht het signaal te reconstrueren, uitgaande van de door de temporele decompositie berekende weegfuncties en doelvectoren. Daarbij stuitte ik op het volgende probleem : bij de bepaling van de amplitudecoëfficiënten wordt de fase van het signaal verwaarloosd, waardoor bij de bepaling van autocorrelatiecoëfficiënten uit het powerspectrum dusdanige afwijkingen ontstaan, dat het beschreven filter instabiel wordt. Voor eventueel later onderzoek geef ik hier de routine die voor de resynthese gebruikt werd, en een uitgebreidere beschrijving van de tests die uitgevoerd zijn.

Als eerste volgt hier de routine die ik geschreven heb om de spectrale amplitudecoëfficiënten om te zetten in filterparameters en op te slaan in een a/p-file.

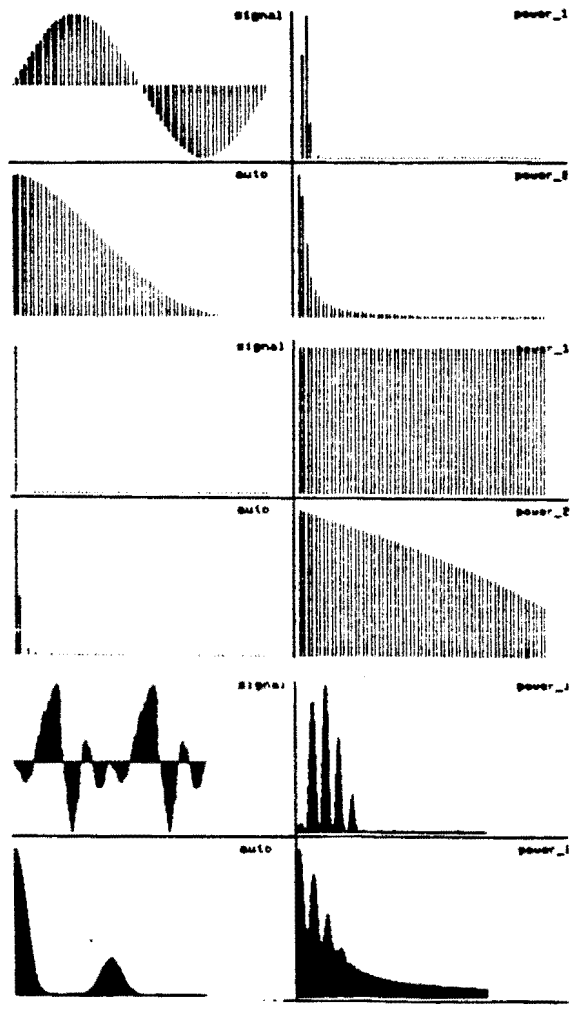
```
c-----------------------------------------------------------------------
c-----------------------------------------------------------------------
c
      subroutine apdump (fn, fw, buf, np, npars, ib,ie)
c
c-----------------------------------------------------------------------
c
c     subroutine to put a buffer of spectrum parameters back into
c     an A/P file.
c     BUF spans the whole file, but contains only valid data from
c     IB to IE
c
c     copyright IPO 5/10/87
c     Lyon Lemmens  for use in editspecsc
c
c-----------------------------------------------------------------------
c
c     description of the parameters :
c
c     fn    ____________ filename of the old file
c     fw    ____________ filename of the new file
c     buf   ____________ buffer containing amplitude coefficients
c                        ( dimension np * (ie-ib) )
c     np    ____________ number of parameters as reported by main
c                        program
c     npars ____________ number of parameters as reported by this
```

```
c                        routine (on exit)
c     ib    ___________ first frame to be converted
c     ie    ___________ last frame to be converted
c
c-----------------------------------------------------------------------
c
c     routines called by this program :
c
c     flexist, flcreate, flopen, |     standard read/write
c     rread, rwrite, rclose,     |-->  LVS-routines for a/p-files,
c     idrex                      |     see IPO Manual no. 68
c
c     fftlvs   : LVS Fast Fourier Transform
c     symvc    : LVS filtersynthesis from autocorrelation coefficients
c     fortr    : LVS transformation of a-parameters in pq-parameters
c     ---------------------------------------------------------------
c                               |
c                               |-->   LVS library subroutines
c
c-----------------------------------------------------------------------
c-----------------------------------------------------------------------
c
      implicit integer*2 (i-n)
c
      character*(*)    fn,fw
      integer*2        is(256)
      integer*4        np, npars, ib,ie, nloop
      double precision buf(np,*), arg
      real               R(128), Ri(128)
      real             y(128,1024), yi(128)
      real             K(128), FC(128)
      logical          first
c
c.....INITIATE Ri AND yi TO ZERO
c
      do i = 1, 128
         Ri(i) = 0.0
         yi(i) = 0.0
      enddo
c
c.....FORCE DOUBLE PRECISION AMPLITUDE TO REAL POWER SPECTUM
c
      do i = 1, npars
         do j = 1, ie
            arg = buf(i,j) * buf(i,j)
            y(i,j) = arg
         enddo
      enddo
c
c.....DETERMINE ORDER OF THE FILTER
c
      nbits=1
      n=1
      do while(2*n.le.np)
         nbits=nbits+1
         n=n*2
      enddo
c
c.....SEE IF FILE EXISTS
```

```
c
      call flexist (%ref(fn),ierr,llp)
      if (ierr.ne.0) print *,'*** APDUMP : file does not exist'
      if (ierr.ne.0) return
c
      call flcreate (%ref(fw),0,0,ierr)
      if (ierr.ne.0) stop '*** APDUMP : cannot create file'
c
      call fopen (%ref(fn),llp,2,1)
      call fopen (%ref(fw),lnall,3,-1)
c
c.....READ IDENTIFICATION RECORDS
c
      call rread (2,is,256)
      nblk=is(35)
      lfr=is(36)
      is(112)=2                ! output are PQ-parameters
      is(113)=npars
      m  = npars
      mh = m/2
      mp = m+1
      mpp = m+2
      m6 = m+6
c
      call rwrite (3,is,256)
      call idrex (2,3,is,is(11))
c
c.....FRAMES TRANSPORTEREN
c
      first = .true.
c
c.....BEGIN VAN DE LUS
c
      do nloop = 1, nblk
         call rread (2,is,lfr)
         if (nloop.ge.ib .and. nloop.le.ie) then
c
c..........TRANSFORM SPECTRAL PARAMETERS TO AUTO-COR-COEFFICIENTS
c
            call fftlvs ( y(1,nloop), yi, R, Ri, m, nbits, 3 )
c
c..........TRANSFORM A-C-COEFFICIENTS TO A-PARMS
c
            call symvc ( R, FC, K, m )
            print '(''+'',i3, A, i3, A)',nloop,' of ',ie,' frames.'
c
c..........CALCULATE PQ-PARAMETERS
c
            call fortr ( FC, is, mp, first, nloop )
            first = .false.
c
c..........FILL IN THE IDENTIFICATION RECORD
c
            do i = m6,17,-1
               is(i+4) = is(i)
            enddo
            do i = 17,20
               is(i) = 0
            enddo
```

```
c
         endif
c
c........WRITE IDENTIFICATION FILE
c
         call rwrite (3,is,lfr)
c
c........END OF THE LOOP
c
      enddo
c
c.....SET THINGS TO ORDER
c
      call rclose (2,ivar)
      call rclose (3,ivar)
c
c.....AND RETURN
c
      return
      end
c
c-----------------------------------------------------------------------
c-----------------------------------------------------------------------
```

Om de procedere te testen ben ik allereerst uitgegaan van een aantal eenvoudige signalen. Ik heb een sinus gegenereerd en hiervan het spectrum bepaald met de fftlvs-routine. De zo verkregen amplitudecoëfficiënten heb aan de routine aangeboden, en na elke stap in de procedure heb ik de variabele `buf(i,j)` laten opslaan. Later heb ik de berekende autocorrelatiecoëfficiënten met een inverse fftlvs weer omgezet in een powerspectrum. Het signaal, eerste powerspectrum, de autocorrelatiecoëfficiënten en het tweede powerspectrum heb ik in één plotje bij elkaar laten zetten.
Dit geheel heb ik nog een keer gedaan, uitgaande van een blok, een puls, en de klinker /ə/. In figuur A.1 ziet u het plotje van de sinus, de puls en de schwa. Er is duidelijk te zien dat het spectrum van het signaal aangetast wordt.



Figuur A.1: Resultaat van het twee maal Fourier transformeren van een signaal, waarbij tussendoor de fase nul gesteld wordt. Power_1 *is het eerst berekende powerspectrum,* Power_2 *is gereconstrueerd uit de autocorrelatiecoëfficiënten* Auto.