Eindhoven University of Technology

MASTER

# Een onderzoek naar de excitatiefunctie van LPC-gesynthetiseerde spraaksignalen m.b.v. pitch-synchrone analyse-resynthese

Benning, F.J.

*Award date:*
1994

Link to publication

**Disclaimer**
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 21. Apr. 2025

Herr (N)
4
2x

Technische Universiteit Eindhoven
Faculteit der Technische Natuurkunde

# Een onderzoek naar de excitatiefunctie van LPC-gesynthetiseerde spraaksignalen m.b.v. pitch-synchrone analyse-resynthese.

F.J. BENNING

oktober 1988

Verslag van een afstudeeronderzoek, verricht in de Akoestisch-Fonetische Groep van het Instituut voor Perceptie Onderzoek te Eindhoven als onderdeel van het doctoraal examen in de Technische Natuurkunde.

Uitgevoerd o.l.v. Prof.Dr. H. Bouma (IPO)
Contactdocent fac. N (TUE) : Prof.Dr. J.A. Poulis
Begeleiding : Ir. J.H. Eggen, Ir. L.F. Willems

## Abstract

During this project it has been investigated to what extent the naturalness of synthetic speech can be improved by using a less simplefied excitation-function than applied in the existing IPO analysis-resynthesis system.

First an excitationfunction of a certain fixed shape has been implemented, based on parametric models as described in literature. Further an excitationfunction based on time varying parameters has been used. The momentary values of these parameters are defined by the natural speechsignal.

In order to apply these excitationfunctions a pitch-synchronous analysis-resynthesis system has been developed.

Using the excitationfunction based on fixed parameters, it turned out to be possible to improve the naturalness for certain speakers. The way in which the variable excitationfunction in this project has been parametrized, didn't yield the intended result. During the development of this function however, modelling of the dynamic behaviour of the excitationfunction seemed to be a potential possibility to improve the naturalness of synthetic speech.

## Samenvatting

In dit afstudeerproject is onderzocht in hoeverre de natuurlijkheid van synthetische spraak verbeterd kan worden door het gebruik van een minder vereenvoudigde excitatiefunctie dan in het bestaande IPO analyse-resynthese systeem wordt toegepast.

Allereerst is een excitatiefunctie geïmplementeerd met een bepaalde vaste vorm op basis van in de literatuur beschreven parametrische modellen. Daarnaast is een excitatiefunctie gebruikt, waarvan bepaalde parameters variabel in de tijd zijn en waarbij de momentane waarden van deze parameters uit het natuurlijke spraaksignaal bepaald worden.

Ten behoeve van het gebruik van deze excitatiefuncties is een pitch-synchroon analyse-resynthese systeem ontwikkeld.

Door toepassing van de excitatiefunctie beschreven door vaste parameters, blijkt de natuurlijkheid voor sommige sprekers verbeterd te kunnen worden. De manier waarop in dit onderzoek de variabele excitatiefunctie geparametriseerd is, blijkt niet het gewenste resultaat op te leveren. Toch bleek bij de ontwikkeling van deze functie, de modellering van het dynamisch gedrag van de excitatiefunctie een potentiële mogelijkheid te bieden tot verbetering van de natuurlijkheid van synthetische spraak.

# Inhoud

# Hoofdstuk 1

# Inleiding

Een van de doelstellingen binnen de Akoestisch-Fonetische Groep van het Instituut voor Perceptie Onderzoek (IPO) is het ontwikkelen van synthetische (kunstmatig opgewekte) spraak, die zich perceptief niet onderscheidt van de natuurlijke spraak van de menselijke stem. Essentiëel voor synthetische spraak is dat aan de produktie ervan een model ten grondslag ligt, waarvan de parameters fysische eigenschappen van het natuurlijke spraaksignaal representeren. Het op het IPO gebruikte systeem voor het genereren van synthetische spraak berust op de methode van analyse-resynthese, waarbij, uitgaande van het zgn. bron-filtermodel, de modelparameters volgens de techniek van lineaire predictie (LPC) rechtstreeks uit het natuurlijke spraaksignaal verkregen worden.

De zo verkregen LPC-gesynthetiseerde spraak wordt gekenmerkt door een goede verstaanbaarheid. De natuurlijkheid van de spraak laat echter te wensen over. In de literatuur (Rosenberg,1971 / Holmes,1973) wordt als een van de mogelijke oorzaken voor het verloren gaan van de natuurlijkheid de excitatiefunctie aangegeven, waarmee het, met het spraakkanaal corresponderende, LPC-filter geëxciteerd wordt. In de huidige LPC-synthese wordt als excitatiefunctie een deltapuls gebruikt.

De doelstelling van dit afstudeerwerk was nu te onderzoeken óf en in hoeverre de natuurlijkheid van de spraak verbeterd kan worden door het gebruik van een andere excitatiefunctie. Het idee hierbij was om i.p.v. de deltapuls allereerst een functie te gebruiken met een bepaalde *vaste* vorm, waarvoor in de literatuur verschillende parametrische modellen genoemd worden (Rosenberg,1971 / Fujisaki & Ljungqvist,1986 en 1987). Daarnaast

was het de bedoeling een functie te implementeren, waarvan bepaalde parameters *variabel* in de tijd zijn, waarbij de waarden van deze parameters uit het natuurlijke spraaksignaal bepaald worden. Tevens was het de bedoeling het effect van het gebruik van deze excitatiefuncties voor meerdere sprekers na te gaan.

Teneinde de excitatiefunctie per pitchperiode te kunnen manipuleren, was het noodzakelijk over een *pitch-synchroon* analyse-resynthese systeem te beschikken, waarbij de plaats en de lengte van het tijdsvenster, waarover de analyse respectievelijk de resynthese wordt uitgevoerd, telkens bepaald worden door de momentane grondtoonperiode. Dit in tegenstelling tot het op het IPO reeds bestaande systeem voor analyse-resynthese van spraak, waarbij een analysevenster met een vaste lengte gebruikt wordt, dat telkens over een vast interval verschoven wordt. De ontwikkeling van een pitch-synchroon analyse-resynthese systeem vormde zodoende een essentiëel onderdeel van het onderzoek.

Tenslotte was het de bedoeling de eventueel gemaakte verbeteringen ten aanzien van de excitatiefunctie perceptief te evalueren d.m.v een experiment, waarin proefpersonen gevraagd wordt de diverse aangeboden synthetische spraakstimuli op hun natuurlijkheid te beoordelen.

De indeling van dit verslag is als volgt. Allereerst wordt in hoofdstuk 2 het bron-filtermodel beschreven, waarop de spraakproduktie gebaseerd is en dus centraal staat in het analyse-resynthese systeem. Vervolgens wordt in hoofdstuk 3 het ontwikkelde pitch-synchrone analyse-resynthese systeem beschreven. Hoofdstuk 4 gaat nader in op de parametrisering van de excitatiefunctie. In hoofdstuk 5 wordt de opzet van het perceptieve experiment beschreven en worden de resultaten hiervan gepresenteerd en geïnterpreteerd. Hoofdstuk 6 bevat tot slot de conclusies van dit afstudeeronderzoek.

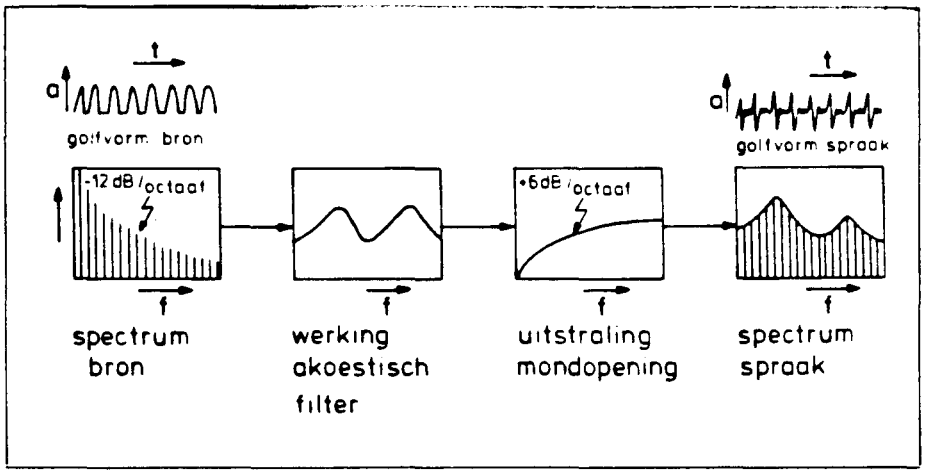
# Hoofdstuk 2

# Bron-filtermodel voor spraakproduktie

## 2.1 Fysica van de spraakproduktie - bron-filtermodel

Een vereenvoudigde voorstelling van de fysica van de spraakproduktie wordt gegeven door het in de experimentele fonetiek algemeen aanvaarde *bron-filtermodel* van Fant (1960). Hierin wordt een *geluidsbron* onderscheiden, waardoor het brongeluid wordt opgewekt, en een *filter*, waardoor het spraakgeluid de gewenste klankkleur (timbre) krijgt.

Het brongeluid wordt gecreëerd doordat een door de longen opgewekte luchtstroom ergens in de mond-keelholte een vernauwing passeert. Treedt deze vernauwing bij de stembanden op, waarbij de stembanden zich periodiek openen en sluiten, dan ontstaat een stemhebbende (periodieke) klank. De periode van de stembandtrilling bepaalt hierbij de waargenomen toonhoogte van het spraakgeluid. De door de stembanden veroorzaakte luchtdrukveranderingen hebben bij benadering een driehoekig verloop als functie van de tijd, waarbij ruwweg de helft van de trillingsperiode de stemspleet geheel gesloten is. De opgewekte geluidsenergie is dus telkens geconcentreerd in een vrij korte tijdsduur, zodat het energiespectrum zich over een groot frequentiegebied uitstrekt. De stembandtrillingen bevatten dus veel harmonischen, waarvan de amplitude in eerste benadering kwadratisch afneemt met de frequentie, zodat de omhullende van het energiespectrum




Figuur 2.1: *Schematische voorstelling van het tot stand komen van een stemhebbende klank volgens de bron-filtertheorie van spraakproduktie (Vogten,1988).*

een helling van ongeveer -12 dB/octaaf heeft.

Behalve bij de stembanden kunnen ook op andere plaatsen in de mondholte zodanige vernauwingen optreden, dat de luchtstroom in krachtige wervelingen wordt gebracht. Hierdoor ontstaan ruisgeluiden (stemloze, niet-periodieke klanken), die eveneens een breed energiespectrum hebben. Door het veranderen van de luchtstroom kan de spreker zowel de amplitude van de stembandtrilling als ook de energie van de ruisgeluiden en daarmee dus de uiteindelijke geluidssterkte van het spraakgeluid regelen.

Het filter representeert nu de akoestische eigenschappen van dat deel van het spraakkanaal dat zich tussen de geluidsbron en de buitenlucht bevindt. Door resonanties in dit kanaal worden bepaalde frequenties door het filter versterkt en andere verzwakt, waardoor een spectrum met een veel grilliger gevormde omhullende ontstaat.

Wanneer de trillende lucht uiteindelijk via de mond en/of neus naar buiten stroomt, treedt in dit laatste stadium nog een stralingseffekt op van de uitstroomopening. De omhullende van het energiespectrum ondergaat hierbij een extra (tijdsonafhankelijke) verandering, overeenkomend met een helling van ongeveer +6 dB/octaaf.

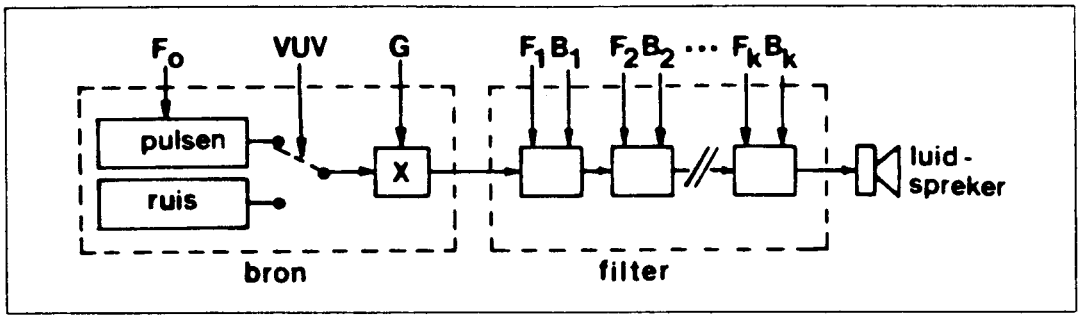
Het uiteindelijke resultaat is het spectrum van een spraakklank, zoals schematisch is weergegeven in figuur 2.1 voor een stemhebbende klank. We zien hierin de frequentie van de stembandtrilling terug in de afstand tussen de harmonischen in het spectrum. Het effekt van de filterende werking van het spraakkanaal uit zich vooral in de plaatsen van de pieken in de spectrale omhullende, de zgn. formanten. Deze plaatsen hangen samen met de vorm van het akoestisch filter en zijn karakteristiek voor met name afzonderlijke klinkers en tweeklanken.

Essentiëel voor het bron-filtermodel is dat bronsignaal, akoestisch filter en stralingseffekt als onafhankelijke elementen beschouwd worden en elkaar niet belasten. Het model beschrijft de fysische eigenschappen van het spraaksignaal als funktie van de tijd door middel van een variabel bronsignaal dat als ingangssignaal dient voor een eveneens variabel lineair filter.

## 2.2 Het toegepaste model voor synthetische spraakproduktie

Het vereenvoudigde bron-filtermodel, waarop de synthetische spraakproduktie in ons analyse-resynthese systeem [1] nu is gebaseerd, is weergegeven in figuur 2.2. Het bronsignaal bestaat voor stemloze spraaksegmenten uit ongecorreleerde witte ruis, terwijl voor stemhebbende stukken spraak een periodieke excitatiepuls met herhalingsfrequentie $F_0$ wordt gebruikt. Welke van de twee bronnen als ingangssignaal voor het filter dient, wordt bepaald door een binaire stemloos/stemhebbend parameter VUV. Verder is als bronparameter in het model een variabele versterkingsfaktor G opgenomen, waarmee de amplitude van het bronsignaal als funktie van de tijd wordt gerepresenteerd. Het akoestisch filter wordt gevormd door een cascade van vijf 2e-orde digitale deelfilters, die ieder één formant beschrijven en gekarakteriseerd worden door een afstemfrequentie F en een bandbreedte B. Het aantal deelfilters bedraagt in principe vijf, aangezien tot 5 kHz bandbegrensde spraak doorgaans niet meer dan vijf formanten bevat. Het filter kan dus afdoende beschreven worden door in totaal 10 filterparameters (in de hier beschreven realisatie : 5 afstemfrequenties en 5 bandbreedten).

[1] Het systeem bestaat in hoofdzaak uit Pascal software-programma's, die uitgevoerd kunnen worden op de IPO-VAX/8530 computer.




Figuur 2.2: *Vereenvoudigd bron-filtermodel voor de produktie van synthetische spraak.* $F_0$ *: herhalingsfrequentie excitatiepulsen, VUV : stemloos/-stemhebbend parameter, G : amplitudeversterkingsfaktor,* $F_k$,$B_k$ *: afstemfrequentie resp. bandbreedte van deelfilter (Vogten, 1988).*

Bron en filter zijn bestuurbaar : door variatie van hun parameters op discrete punten in de tijd kan de tijdsontwikkeling van het spraakgeluid beschreven worden. De momentane waarden van de modelparameters worden hierbij verkregen via een analyse van het originele spraaksignaal.

In vergelijking met de fysica van de menselijke spraakproduktie is het hier beschreven produktiemodel een sterke vereenvoudiging en vertoont door zijn beperkingen en eenvoud ook duidelijke verschillen. Zo voorziet het model niet in een gelijktijdige kombinatie van periodiek- en ruisbron (wat bijvoorbeeld bij stemhebbende wrijfklanken voorkomt) en evenmin in interaktie-effekten tussen bron en filter. Verder wordt het spraaksignaal in het model beschreven als een stapsgewijze opeenvolging van stationaire signalen, terwijl de modelparameters in feite continu variëren. Toch is het model goed bruikbaar omdat juist door deze beperkingen en eenvoud de modelparameters snel en automatisch rechtstreeks uit het originele spraaksignaal bepaald kunnen worden. Hoe deze analyse in z'n werk gaat, zal nu in het volgende hoofdstuk beschreven worden.

# Hoofdstuk 3

# Pitch-synchrone analyse en resynthese

## 3.1 Inleiding

In het vorige hoofdstuk is een model beschreven, waarmee synthetische spraak geproduceerd kan worden. In dit hoofdstuk zal nu uiteengezet worden hoe de parameters van dit model als funktie van de tijd uit het originele spraaksignaal verkregen kunnen worden (de analyse) en hoe vervolgens uit deze stuurgetallen weer spraaksignalen gevormd worden (de resynthese). De analyse vindt steeds plaats over een relatief kort tijdsinterval, het zgn. *analysevenster*, waarbinnen de modelparameters constant verondersteld worden.

Centraal in dit onderzoek staat de excitatiefunktie, waarmee het filter in het geval van stemhebbende spraaksegmenten periodiek met herhalingsfrequentie $F_0$ geëxciteerd wordt. Willen we deze excitatiefunktie kunnen manipuleren, dan is het noodzakelijk over een analyse-resynthese systeem te beschikken, waarbij zowel de plaats als de lengte van het analysevenster synchroon gekozen worden aan de momentane grondtoonperiode, d.i. het tijdsinterval waar binnen zich één excitatiepuls voordoet (in het vervolg *pitch-periode* genoemd). Het op het IPO reeds bestaande analyse-resynthese systeem (LVS) (Vogten,1983) is echter gebaseerd op een analysevenster met een *vaste* lengte (25ms), dat telkens over een *vaste* vaste afstand (10ms) in de tijd verschoven wordt. Zodoende was het noodzakelijk een *pitch-synchroon* analyse-resynthese systeem te ontwikkelen. Het

principe en de uitvoering hiervan zal nu in dit hoofdstuk behandeld worden.

## 3.2 Het bepalen van de pitch-periodes

Om een spraakfragment pitch-synchroon te kunnen analyseren resp. resynthetiseren, zullen allereerst zowel de plaats als de lengte van de pitch-periodes in het te analyseren signaal bekend moeten zijn. Wanneer men de golfvorm van een stemhebbend spraakfragment bekijkt (zie bv. figuur 2.1 rechtsboven), zou piekdetectie, waarbij de maxima in het spraaksignaal als indicatie voor de pitch-periodes gebruikt worden, een voor de hand liggende methode lijken. Deze methode levert echter spoedig problemen, indien het signaal minder duidelijk gepiekt is en is bovendien gevoelig voor polariteitsveranderingen (fasedraaiingen) van het signaal.

Een bevredigend werkende methode ter bepaling van de pitch-periodes is ontwikkeld door Eggen (IPO). Deze methode gaat uit van een LPC-analyse [1] van het spraaksignaal, waarbij telkens over een zeer klein venster ( $\pm 3$ msec ) de parameters van het filter m.b.v. de covariantiemethode (Markel & Gray,1976) bepaald worden. De mate waarin dit optimaliseringsproces faalt, wordt weergegeven door een restfout, de zgn. covariantiefout. Door nu het korte venster telkens één sample in het spraaksignaal op te schuiven en opnieuw een analyse uit te voeren, wordt uiteindelijk het verloop van de covariantiefout als functie van de tijd verkregen. Er kan aangetoond worden, dat de covariantiefout groot wordt op plaatsen in het spraaksignaal, waar de maximale excitaties plaatsvinden. Het verloop van de covariantiefout vormt zodoende een indicatie voor het bepalen van de pitch-periodes.

Hiervan uitgaande is door Eggen een algorithme geïmplementeerd, dat lokaal zoekt naar maxima in de covariantiefout. Deze maxima worden vervolgens gebruikt om de pitchperiodes aan te geven. Uit nader onderzoek met synthetische spraaksegmenten, waar de plaatsen van de excitatiepulsen bekend zijn, is namelijk gebleken dat het begin van een pitch-periode een systematische afstand, bij benadering gelijk aan de orde van de analyse, vóór het maximum in de covariantiefout ligt. Het begin van een pitch-periode wordt nu uiteindelijk gevonden door de dichtstbijzijnde nuldoor-

[1] LPC-analyse is een mathematische optimaliseringstechniek en zal in de volgende paragraaf nader beschreven worden (dan echter gebaseerd op de autocorrelatiemethode).

gang van het spraaksignaal rond deze positie te bepalen.

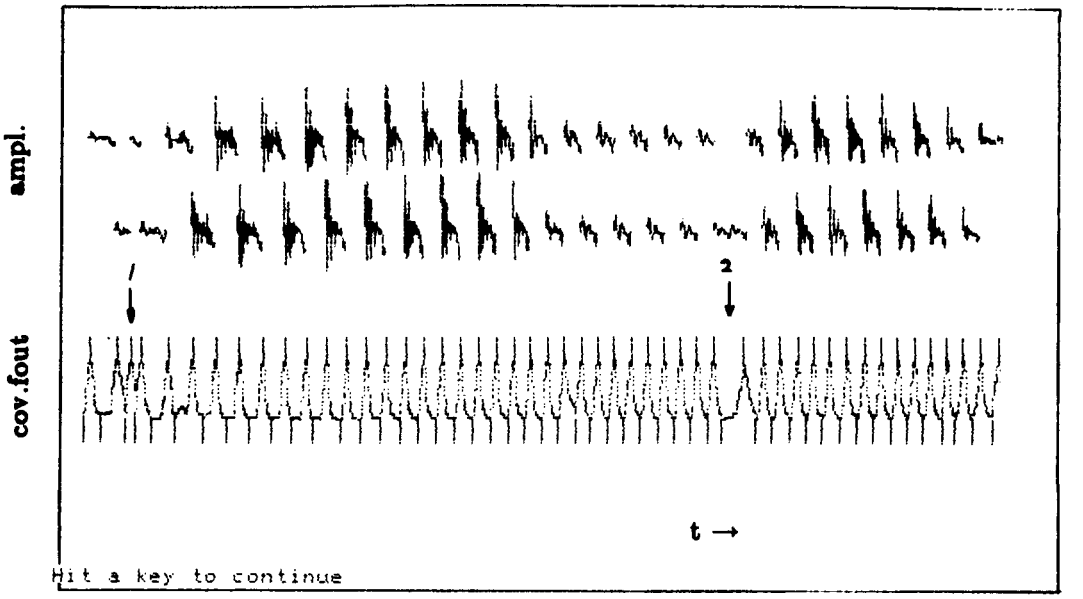
Bovenbeschreven methode levert zodoende de pitch-periodes voor *stemhebbende* spraaksegmenten. Aangezien in *stemloze* spraaksegmenten per definitie geen grondtoon aanwezig is, wordt hier een alternatieve keuze voor het analysevenster gedaan in de vorm van een *vast* venster ter lengte van 100 samples (=10ms), wat overeenkomt met de gemiddelde lengte van een pitch-periode in een mannenstem met een toonhoogte van $\pm$100 Hz.

Er zal dus bij de analyse onderscheid gemaakt moeten worden tussen stemloze en stemhebbende spraaksegmenten, weergegeven door de binaire parameter VUV. Deze parameter wordt verkregen middels het LVS-programma PCT, dat een toonhoogtemeting en daarmee tevens een stemloos/stemhebbend beslissing uitvoert (Hermes,1986). De bij deze stemloos/stemhebbend detectie eventueel gemaakte fouten kunnen nog m.b.v. het LVS-programma CHP [2] gecorrigeerd worden (Dit bleek bij de in dit onderzoek gebruikte stimuli overigens nauwelijks nodig).

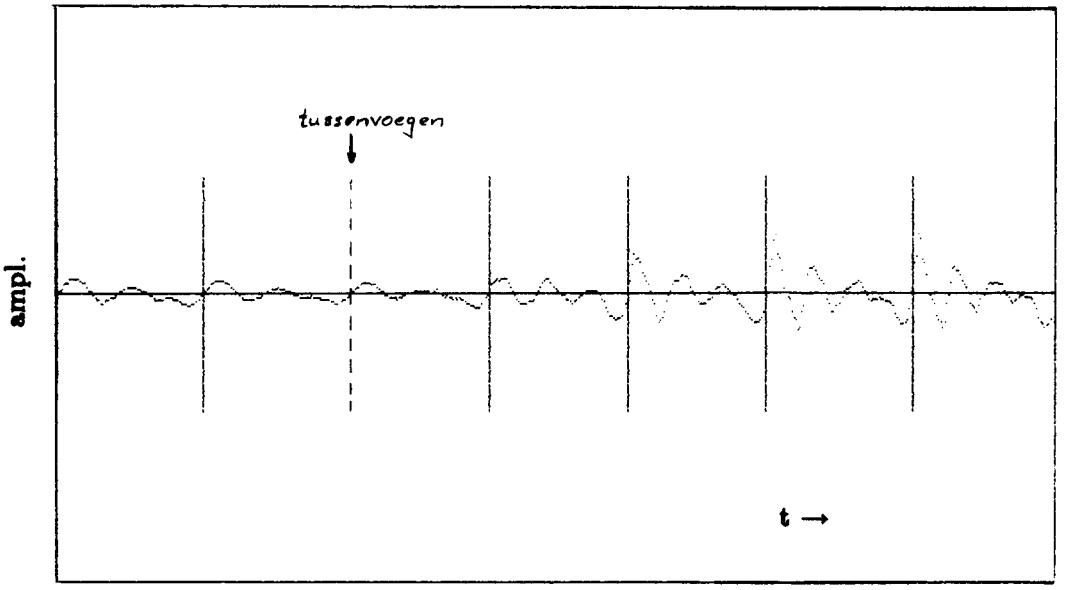
Een voorbeeld van het resultaat van een 'pitch-analyse' van een stemhebbend spraakfragment is weergegeven in figuur 3.1. Hierin is het verloop van de covariantiefout als functie van de tijd onderin weergegeven. De vertikale streepjes aan de boven- en onderkant van het signaal representeren de gevonden maxima respectievelijk minima. Boven in de figuur zijn nu (om en om) de pitch-periodes weergegeven, zoals deze volgens bovenbeschreven algorithme uit het geanalyseerde spraaksignaal gesneden zijn. Het merendeel van de periodes blijkt correct uitgesneden te worden. In een aantal gevallen, met name op plaatsen waar het spraaksignaal klein is en weinig periodieke structuur vertoont én op plaatsen die overeenkomen met overgangen tussen stemhebbende en stemloze spraakgedeelten, levert de methode verkeerde resultaten. Zo geeft pijl no.1 links in figuur 3.1 een ten onrechte gevonden beginpunt van een periode aan, terwijl pijl no.2 rechts in de figuur een niet-gevonden pitch-periode aanduidt (octaaffout).

Daarom is ter aanvulling van bovenbeschreven 'pitch-analyse' nog een interactief edit-programma ontwikkeld, waarmee de gevonden pitch-periodes gecontroleerd en zonodig gecorrigeerd kunnen worden. Hiertoe worden de gevonden pitch-periodes d.m.v. markeringen grafisch in de (sterk vergrootte ) golfvorm van het geanalyseerde spraaksignaal weergegeven.

[2]Het LVS-systeem omvat naast programmatuur voor de analyse-resynthese van spraak tevens de mogelijkheid tot het interactief manipuleren van de analyseparameters (Vogten,1985).




Figuur 3.1: *Voorbeeld van het resultaat van een 'pitch-analyse' ter bepaling van de pitch-periodes in het stemhebbende spraakfragment "maanda" (uit de zin : "Maandag gaan we naar het zwembad").*




Figuur 3.2: *Voorbeeld van het 'edit'-proces voor het controleren respectievelijk corrigeren van gevonden pitch-periodes. D.m.v. vier opties kunnen de beginpunten van de pitch-periodes achtereenvolgens gecorrigeerd worden (zie tekst).*

De gebruiker kan vervolgens d.m.v. een aantal opties ieder gevonden beginpunt van een periode goedkeuren, verschuiven (over een op te geven afstand), verwijderen of een beginpunt van een periode tussenvoegen. (de correcties bleken overigens voor het merendeel neer te komen op het corrigeren van octaaffouten). Het aantal pitch-periodes dat op deze manier gecorrigeerd moet worden, blijkt over het algemeen hooguit 5% te bedragen. Figuur 3.2 geeft een voorbeeld van het 'edit'-proces voor een segment (rond pijl no.2) uit het spraaksignaal van figuur 3.1.

## 3.3 LPC-analyse en resynthese

In hoofdstuk 2 is het bron-filter model voor spraakproduktie besproken. Hierbij onderscheiden we (in het geval van stemhebbende klanken) een bronsignaal met een spectrale omhullende van -12 dB/octaaf, een met het spraakkanaal corresponderend filter met een vlakke omhullende en een uitstralingseffekt van +6 dB/octaaf (zie figuur 2.1). Aangezien deze drie componenten lineair en onafhankelijk verondersteld werden, mogen we ze als een geheel opvatten, wat één filter oplevert met een helling van -6 dB/octaaf. Dit filter (verder *synthesefilter* genoemd) bepaalt dan de omzetting van het vlakke bronspectrum naar het gekleurde, gepiekte spectrum van de spraak. Het feit dat het bronspectrum nu vlak is, levert de mogelijkheid om via de techniek van 'invers filteren' of Linear Predictive Coding (Markel & Gray,1976) de parameters van het synthesefilter rechtstreeks uit het spraaksignaal te bepalen. Weten we namelijk een spraaksegment met een *analysefilter* zodanig te filteren, dat het uitgangssignaal van dit filter een vlak spectrum heeft, dan moet de overdrachtsfunctie van dat analysefilter de geïnverteerde zijn van het synthese(produktie-)filter dat we zoeken, immers dit laatste heeft een vlak ingangsspectrum. De parameters van het synthesefilter kunnen dus gevonden worden door de overdrachtsfunktie van het analysefilter te inverteren. Hoe de parameters van het analysefilter zó berekend worden, dat het uitgangsspectrum vlak wordt, zal nu in het kort uiteengezet worden. Voor een meer uitgebreide behandeling van de LPC-techniek wordt de lezer verwezen naar de literatuur op dit gebied (Atal & Hanauer,1971 / Makhoul,1975 / Markel & Gray,1976).

### 3.3.1 Bepaling van de $a$-parameters van het analysefilter

Bij de LPC-analyse wordt het analysefilter voorgesteld door een digitaal recursief filter van de orde $M$, waarvan het uitgangssignaal $e_n$ (sample op tijdstip $n/f_s$ met $f_s$ de bemonsteringsfrequentie en n een geheel getal) gegeven wordt door de som van het ingangssample $s_n$ op datzelfde tijdstip en een lineaire combinatie van $M$ voorgaande ingangssamples :

$$e_n = s_n + \sum_{k=1}^{M} a_k s_{n-k} = \sum_{k=0}^{M} a_k s_{n-k} \qquad (a_0 := 1) \tag{3.1}$$

De orde $M$ van het filter geeft het aantal voorgaande samples aan, ieder voorzien van een weegfactor $a_k$, dat bijdraagt tot het uitgangssample $e_n$. Voor de totale energie van het uitgangssignaal gedefiniëerd door :

$$E = \sum_{n} e_n^2 \qquad n = -\infty, \ldots, \infty \tag{3.2}$$

kan uitgaande van (3.1) afgeleid worden (Vogten,1983) dat deze minimaal is indien geldt :

$$\sum_{k=1}^{M} a_k R_{i-k} = -R_i \qquad i = 1, \ldots, M \tag{3.3}$$

waarin

$$R_{i-k} = \sum_{n} s_{n-i} s_{n-k} \tag{3.4}$$

de $(i-k)^e$ autocorrelatie van het ingangssignaal $s_n$ definiëert. Hoewel $n$ hierbij in principe van $-\infty$ tot $+\infty$ loopt, zijn bij de analyse per definitie alle ingangssamples $s_n$ buiten het analysevenster, dat uit $N$ samples bestaat, nul. Het stelsel (3.3) van $M$ vergelijkingen met de $M$ filtercoëfficiënten $a_k$ als onbekenden kan recursief worden opgelost na berekening van de autocorrelaties $R_{i-k}$ volgens (3.4).

Wanneer de filtercoëfficiënten voldoen aan (3.3) is de energie $E$ van het uitgangssignaal, het *residusignaal* genoemd, minimaal en deze wordt dan gegeven door :

$$E_m = \sum_{i=0}^{M} a_i R_i \tag{3.5}$$

De kwadratensom (3.2) in het tijddomein is nu ook te schrijven als een integratie in het frequentiedomein (theorema van Parseval) :

$$E = \sum_n e_n^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(\omega)|^2 d\omega \tag{3.6}$$

Minimaliseren van $E$ volgens (3.3) betekent dus dat de spectrale omhullende van het uitgangssignaal dan zo vlak mogelijk is gemaakt, waarmee de filtercoëfficiënten $a_k$ van het analysefilter gevonden zijn.

Het minimaliseren van de energie van het residusignaal $e_n$ gegeven door (3.1) is ook op te vattten als het minimaliseren van de fout, die gemaakt wordt als voor ieder sample $s_n$ een 'voorspelling' $\hat{s}_n$ gemaakt wordt, gedefiniëerd door een lineaire combinatie van $M$ voorgaande samples :

$$\hat{s}_n = \sum_{k=1}^{M} a'_k s_{n-k} \tag{3.7}$$

De bij deze 'lineaire predictie' optredende fout is dan het verschil tussen het werkelijke signaal $s_n$ en de voorspelde waarde $\hat{s}_n$ :

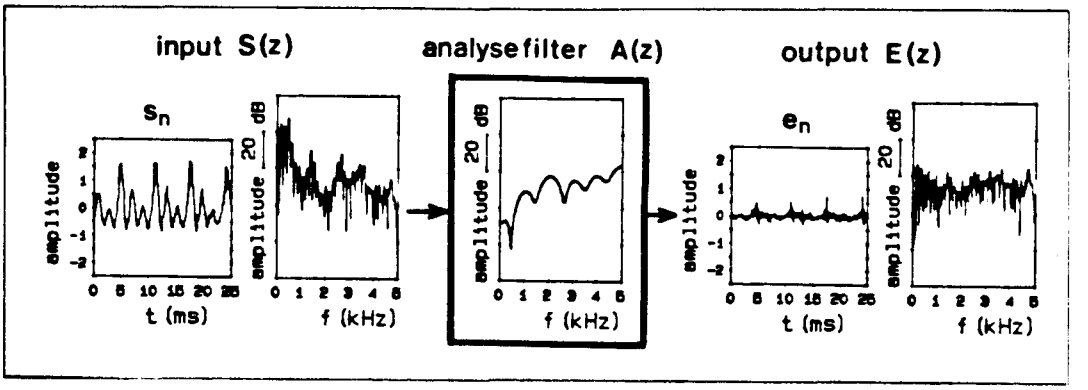
$$e'_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{M} a'_k s_{n-k} \tag{3.8}$$

Minimaliseren van dit verschil $e'_n$ over alle samples binnen het analysevenster levert eenzelfde stelsel vergelijkingen als (3.3), nu echter met als variabelen $a'_k$, die op een min-teken na gelijk zijn aan de coëfficiënten $a_k$. Het teken van de filtercoëfficiënten is echter zuiver een kwestie van definitie.

Een voorbeeld van een $10^e$-orde analysefilter, berekend voor een periodiek ingangssignaal, is weergegeven in figuur 3.3. Links staan golfvorm en spectrum van het ingangssignaal, rechts die van het uitgangs(residu)signaal. Het spectrum van het analysefilter is in het midden weergegeven. Duidelijk is te zien dat het spectrum van het residusignaal is vlakgestreken, doordat de pieken (resonanties) in de omhullende van het ingangsspectrum door dalen (antiresonanties) van het analysefilter 'geneutraliseerd' worden.

### 3.3.2 Het LPC-synthesefilter

Het gezochte synthesefilter kan nu eenvoudigweg verkregen worden door de overdrachtsfunctie van het berekende analysefilter te inverteren. Zouden we




Figuur 3.3: *Voorbeeld van het energiespectrum(midden) van een $10^e$-orde analysefilter $A(z)$, berekend voor een stemhebbend ingangssignaal.*

als ingangssignaal voor dit synthesefilter het residusignaal $e_n$ (uitgangssignaal van analysefilter) nemen, dan verkrijgen we als output van het synthesefilter weer het originele spraaksignaal dat als ingangssignaal voor het analysefilter diende, immers analyse- en synthesefilter zijn elkaars geïnverteerden.

Volgens het in hoofdstuk 2 beschreven produktiemodel voor synthetische spraak, wordt echter als ingangssignaal voor het synthesefilter een sterk 'geïdealiseerd' restsignaal genomen, namelijk het bronsignaal $u_n$ (excitatiepulsen of ongecorreleerde ruis) met amplitudefactor $G$. Het spectrum van het bronsignaal moet hierbij, overeenkomstig het spectrum van het residusignaal, een vlakke omhullende hebben. Het uitgangssignaal $s'_n$ van het synthesefilter, dat het oorspronkelijke spraaksignaal $s_n$ zo goed mogelijk benadert, wordt dan gegeven door :

$$s'_n = Gu_n - \sum_{k=1}^{M} a_k s'_{n-k} \tag{3.9}$$

waarbij de weegfactoren $a_k$ de filterparameters van het berekende analysefilter zijn.

De parameter, die nu nog gespecificeerd zal moeten worden, is de amplitudeversterkingsfactor $G$, waarmee in het model de energie van het spraaksegment geregeld kan worden. Deze parameter wordt bepaald door de eis dat de energie van het gesynthetiseerde spraaksignaal gelijk moet zijn aan die van de oorspronkelijke, geanalyseerde spraak. Bij de beschrijving van de pitch-synchrone synthese (§3.5) zal nader uitgewerkt worden, hoe de $G$-factor uit de energie van het residusignaal bepaald kan worden.

## 3.4 Pitch-synchrone analyse

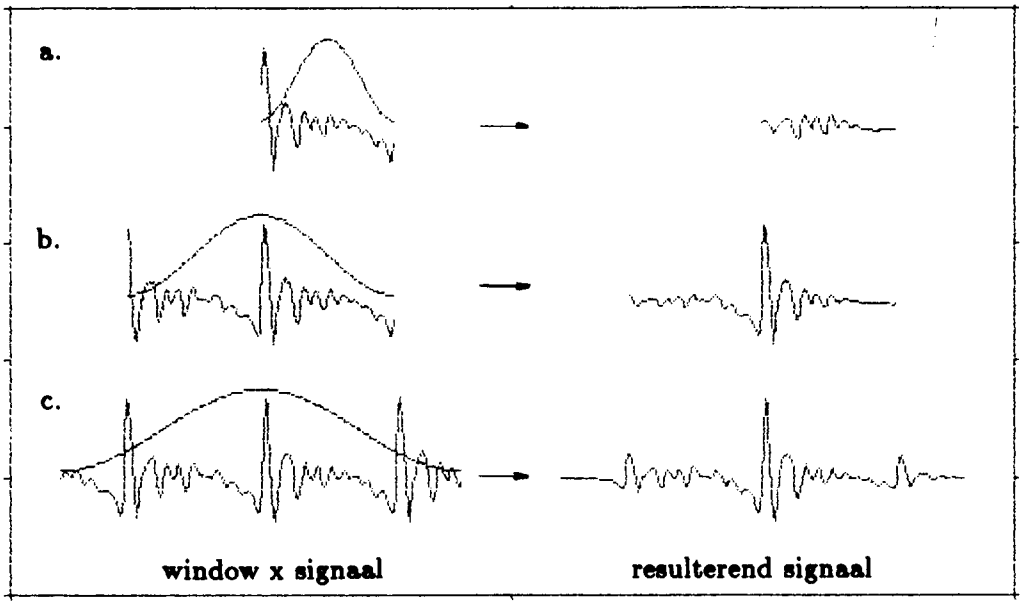
In deze paragraaf zal nu beschreven worden hoe, uitgaande van de in de vorige paragrafen besproken principes voor het bepalen van de pitch-periodes en het berekenen van het LPC-filter, het pitch-synchrone analyse-proces in z'n werk gaat.

Voorafgaande aan de analyse moet het te analyseren spraaksignaal (opgeslagen op magneetband) allereerst in digitale vorm omgezet worden. Hierbij worden alle frequentiecomponenten van het signaal boven de 5 kHz weggefilterd, aangezien deze nauwelijks bijdragen tot de perceptieve kwaliteit van de spraak. Hieruit volgt dan (bemonsteringstheorema van Nyquist) een samplefrequentie van $f_s = 10$ kHz. Het gedigitaliseerde spraaksignaal wordt vervolgens in de computer opgeslagen in een zgn. N-spraakfile.

In het spraaksignaal worden nu eerst de pitch-periodes volgens de in §3.2 beschreven methode bepaald. De beginpunten van de gevonden pitch-periodes worden weggeschreven naar een outputfile. De stemloze 'periodes' worden hierbij ter identificatie van een minteken voorzien. Voor een mannenstem ($F_0 \simeq 125$ Hz) bedraagt de gemiddelde lengte van een pitch-periode $\pm$ 80 samples, voor een vrouwenstem ($F_0 \simeq 300$ Hz) $\pm$ 35 samples.

De eigenlijke pitch-synchrone analyse bestaat nu hierin, dat achtereenvolgens voor elke pitch-periode de parameters van het in de vorige paragraaf besproken analysefilter bepaald worden. Pauzes en andere stukken, waar het signaal (vrijwel) nul is, hoeven echter niet te worden geanalyseerd. Daarom wordt eerst getest of de gemiddelde energie van het signaal binnen het analysevenster een bepaalde drempelwaarde overschrijdt. Alleen als hieraan voldaan is, wordt een analyse uitgevoerd.

De in de vorige paragraaf beschreven LPC-techniek ter bepaling van de filterparameters berustte op het minimaliseren van de energie van het residusignaal over een oneindig groot interval ($-\infty < n < \infty$). In de praktijk is het te analyseren signaal echter slechts over een eindig interval, nl. het analysevenster, bekend. Dit komt overeen met het vermenigvuldigen van het totale signaal met een rechthoekig venster. Een dergelijk ge'window'ed signaal vertoont echter discontinuïteiten aan de randen van het venster, wat tot het foutief bepalen van het filterspectrum zou leiden. Daarom worden de samples binnen het analysevenster vermenigvuldigd met een meer




Figuur 3.4: *Illustratie van het effect van de toepassing van een Hamming-window bij verschillende keuzen van het analysevenster. a) lengte analysevenster = één pitch-periode b) lengte analysevenster = twee pitch-periodes c) lengte analysevenster = drie pitch-periodes.*

vloeiend verlopende windowfunctie, namelijk de zgn. Hamming-windowfunctie (Witten, 1982), gegeven door :

$$H_n = 0.54 - 0.46\cos(2\pi n/N) \qquad n = 1, \ldots, N \qquad (3.10)$$

waarbij $N$ het totaal aantal samples binnen het analysevenster bedraagt. In figuur 3.4a is de toepassing van het window op een stuk signaal ter lengte van één pitch-periode weergegeven.

### Keuze van het analysevenster

Door het toepassen van dit window ontstaat nu echter een probleem t.a.v. de keuze van de lengte van het analysevenster. Kiezen we deze namelijk gelijk aan de lengte van de momentane pitch-periode, dan is in figuur 3.4a te zien, dat het gedeelte van de pitch-periode, dat de meeste informatie bevat, nu juist door het toegepaste window weggedrukt wordt. Daarom

wordt als lengte van het analysevenster een veelvoud van de lengte van de te analyseren pitch-periode genomen. Bij het ontwikkelen van het systeem bleek de precieze keuze van deze lengte alsmede de ligging van het analysevenster echter tamelijk kritisch. Zo bleek een analysevenster gelijk aan *driemaal* de lengte van de te analyseren pitch-periode en symmetrisch gelegen rond het beginpunt van deze pitch-periode (zie figuur 3.4c), tot een onjuiste analyse te leiden. Dit uitte zich bij de resynthese in een onjuist verloop van de amplitudeversterkingsfactor $G$, wat perceptief als een sterke 'krakerigheid' (Eng: 'shimmer') in de gesynthetiseerde spraak waar te nemen was. Uit nader onderzoek bleek dat een juiste analyse verkregen wordt, indien het analysevenster gelijk gekozen wordt aan *tweemaal* de lengte van de te analyseren pitch-periode, zoals is weergegeven in figuur 3.4b. Deze keuze kan aannemelijk gemaakt worden door bestudering van figuur 3.4. Alleen in het geval van figuur 3.4b levert toepassing van het Hamming-window een resulterend analysesignaal, dat overeenkomt met de te analyseren pitch-periode.

De gemiddelde grootte van het analysevenster bedraagt zodoende voor een mannenstem $\pm$ 160 samples en voor een vrouwenstem $\pm$ 70 samples. Zodoende wordt telkens een spraaksegment van $\pm$ 16 respectievelijk $\pm$ 7 msec. ingelezen en geanalyseerd.

### Pre-emphase

Bij de bespreking van de LPC-techniek (§ 3.3) hebben we geconstateerd dat voor stemhebbende spraakfragmenten de globale helling van het langetermijn spectrum ongeveer -6 dB/octaaf bedraagt, t.g.v het gezamelijk effect van bronspectrum (-12 dB/octaaf) en uitstralingseffect (+6 dB/octaaf). Voor stemloze fragmenten vervalt de eerste, zodat dan een gemiddelde helling van +6 dB/octaaf ontstaat. Over langere spraakuitingen zijn stemloze gedeelten echter zowel wat betreft hun duur als amplitude doorgaans sterk in de minderheid.

Daarom wordt vóór de analyse een zgn. *pre-emphasefilter* toegepast, waarvan de overdrachtsfunctie bij benadering een helling van +6 dB/octaaf heeft ter compensatie van bovengenoemde -6 dB/octaaf-helling. Door deze bewerking kan het analysefilter nu nauwkeuriger bepaald worden, immers zonder pre-emphase zou de overall helling van -6 dB/octaaf ook door het analysefilter geneutraliseerd moeten worden.

Het effect van de +6 dB/octaaf-compensatie wordt digitaal verkregen (Witten,1982) door ieder sample $s_n$ van het te analyseren signaal te vervangen door $s'_n$ met :

$$s'_n = s_n + ps_{n-1} \tag{3.11}$$

met de pre-emphaseconstante $p = -0.9$ gekozen in navolging van Vogten (1983).

### Filterberekening

Over het aldus voorbewerkte signaal wordt nu een $10^e$-orde filteranalyse uitgevoerd, zoals beschreven in § 3.3.1. Dit levert als resultaat een set van 10 filtercoëfficiënten, de $a$-parameters, alsmede de energie van het residusignaal binnen het analysevenster. Hiermee is dan de analyse van de betreffende pitch-periode voltooid en worden de $a$-parameters en de energie van het residusignaal tesamen met het beginpunt van de betreffende pitch-periode, eventueel voorzien van een stemloos indicatie, weggeschreven naar een outputfile (de energie van het residusignaal wordt bij de synthese gebruikt ter bepaling van de $G$-factor). Daarmee is dus één pitch-periode geanalyseerd. Vervolgens worden begin- en eindpunt van de volgende pitch-periode ingelezen en wordt de gehele cyclus herhaald. Aldus wordt de gehele spraakuiting doorlopen.

## 3.5 Pitch-synchrone synthese

Bij de pitch-synchrone synthese wordt, uitgaande van de pitch-synchroon verkregen analyseparameters, het spraaksignaal weer per periode geresynthetiseerd door in principe de bij de analyse gevolgde procedure in omgekeerde richting te doorlopen. Als basis voor de resynthese dient hierbij het in § 2.2 besproken bron-filter produktiemodel.

Allereerst worden de analyseparameters (begin- en eindpunt van pitchperiode, stemloos/stemhebbend parameter, energie residusignaal en 10 filterparameters), behorend bij de te synthetiseren pitch-periode, van de analyse-outputfile ingelezen. Het te synthetiseren signaal (ter lengte van de betreffende periode) wordt nu volgens vgl. (3.9) gegeven door :

$$s'_n = Gu_n - \sum_{k=1}^{10} a_k s'_{n-k}$$

waarin $a_k$ de coëfficiënten van het berekende $10^e$-orde analysefilter zijn, $G$ de amplitudeversterkingsfactor is en $u_n$ het ingangssignaal voor het synthesefilter : een periodieke excitatiepuls of witte ruis.

### Het bronsignaal $u_n$

Welke van deze twee bronsignalen wordt gebruikt, wordt bepaald door de stemloos/stemhebbend parameter. Voor stemloze periodes wordt $u_n$ verkregen door een trekking uit randomgetallen tussen $-\frac{1}{2}$ en $+\frac{1}{2}$ (Vogten, 1983). Tot nu toe hebben we de excitatiepuls, waarmee het filter in het geval van stemhebbende periodes geëxciteerd wordt, nog niet nader gespecificeerd. Het doel van dit onderzoek is immers het manipuleren van deze excitatiefunctie. Het volgende hoofdstuk zal hier dan ook uitgebreid over handelen. Op dit moment beperken we ons vooralsnog tot de ontwikkeling van een pitch-synchroon analyse-resynthese systeem. Daarom gebruiken we allereerst als excitatiepuls eenzelfde functie als in het LVS-systeem, namelijk een eenheidsimpuls (deltapuls) met herhalingsfrequentie $F_0$. Voor stemhebbende periodes wordt zodoende $u_n = 1$ genomen als $n$ overeenkomt met het begin van de periode ($1^e$ sample) en elders 0.

### De amplitudeversterkingsfactor G

Zoals reeds bij de bespreking van het LPC-synthesefilter vermeld is, wordt de amplitudeversterkingsfactor $G$ bepaald door de eis dat de energie van het gesynthetiseerde spraaksignaal gelijk moet zijn aan die van de oorspronkelijke, geanalyseerde spraak. Dit betekent dus dat de energie van de excitatie overeen moet komen met de energie van het residusignaal. Dit geeft ons de mogelijkheid om $G$ te bepalen door de gemiddelde energie van de excitatie gelijk te stellen aan de gemiddelde energie van het residusignaal (van Hemert,1987). De totale energie van het residusignaal wordt gegeven door vgl. (3.2) :

$$E_{res} = \sum_{n} e_n^2$$

waarbij $n$ gelijk is aan het totaal aantal samples $N$ binnen het analysevenster. De gemiddelde energie per sample wordt nu verkregen door vgl. (3.2) te delen door het effectief aantal samples $N'$ binnen het analysevenster. Ten gevolge van de toepassing van een windowfunctie bij de analyse (zie § 3.4) is dit effectieve aantal echter niet gelijk aan het totaal aantal samples $N$ binnen het analysevenster, maar geldt :

$$N' = \sum_{n=1}^{N} H_n \tag{3.12}$$

waarbij $H_n$ gegeven wordt door de Hamming-windowfunctie (3.10). Voor $N \gg 1$ gaat (3.12) dan over in :

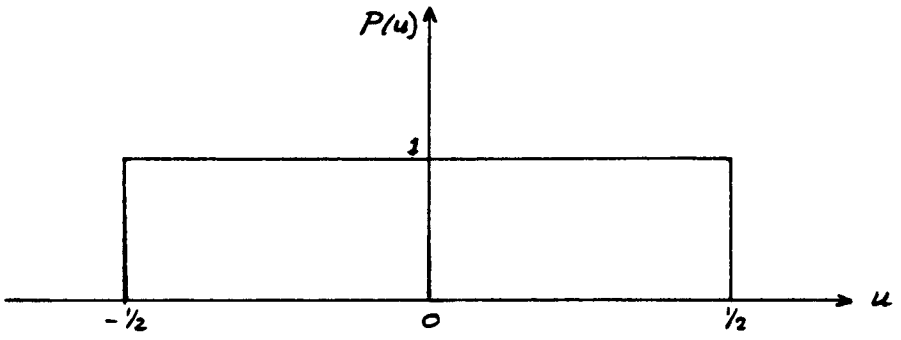
$$N' = 0.54N \qquad N \gg 1 \tag{3.13}$$

zodat de gemiddelde energie van het residusignaal $e_n$ gegeven wordt door :

$$\overline{e_n^2} = \frac{E_{res}}{N'} = \frac{E_{res}}{0.54N} \tag{3.14}$$

Voor stemhebbende periodes geldt nu dat er iedere $T_0$ samples ($T_0$ : lengte periode) een impuls met hoogte $G_v$ gegeven wordt, zodat de gemiddelde energie van het stemhebbende excitatiesignaal $x_n$ gegeven wordt door :

$$\overline{x_n^2} = \frac{G_v^2}{T_0} \tag{3.15}$$




Figuur 3.5: *De kansdichtheidsfunctie $P(u)$ voor $u$.*

Gelijkstelling van (3.15) met (3.14) levert dan de stemhebbende amplitudefactor $G_v$ :

$$G_v = \sqrt{\frac{E_{res} T_0}{N'}} \tag{3.16}$$

Voor stemloze periodes wordt een random signaal $u$ gegenereerd, uniform verdeeld over het interval $[-\frac{1}{2},+\frac{1}{2}]$ (zie figuur 3.5). De gemiddelde energie van het bronsignaal $u$ kan nu worden verkregen door de kansdichtheidsfunctie $P(u)$ te integreren :

$$\overline{u^2} = \int_{-1/2}^{1/2} u^2 P(u) du = \frac{1}{12} \tag{3.17}$$

De gemiddelde energie van het stemloze excitatiesignaal $x$ is dus :

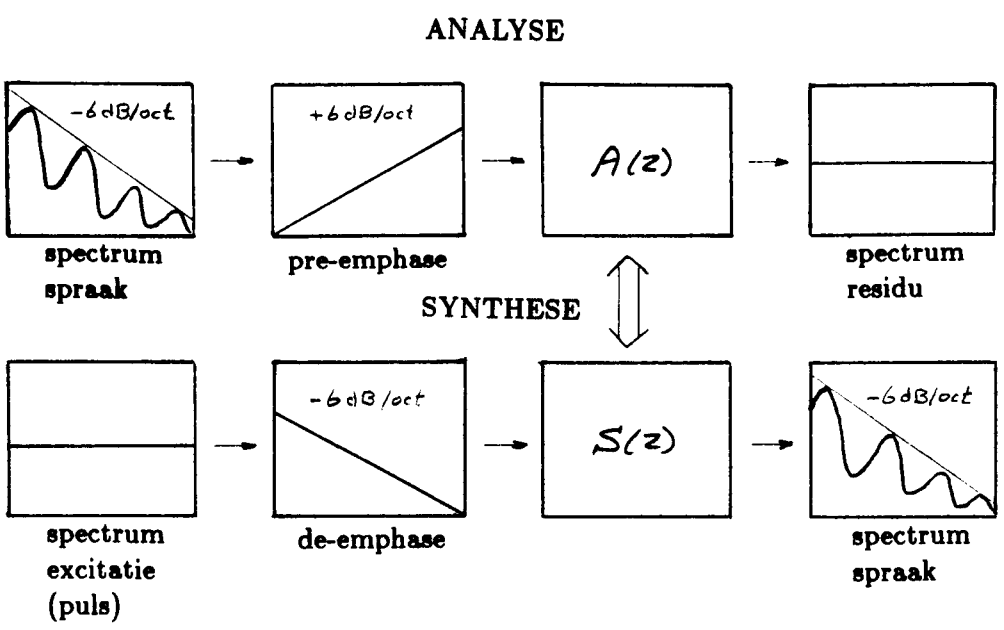
$$\overline{x^2} = \overline{(G_u u)^2} = \frac{G_u^2}{12} \tag{3.18}$$

Door gelijkstelling van (3.18) met (3.14) volgt nu voor de stemloze amplitudefactor $G_u$ :

$$G_u = \sqrt{\frac{12 E_{res}}{N'}} \tag{3.19}$$

### De-emphase

Aangezien bij de analyse een vast pre-emphase filter aan de berekening van het $10^e$-orde analysefilter voorafging, moet nu bij de synthese een omgekeerde filtering (zgn. *de-emphase*) plaatsvinden. Dit de-emphase filter heeft dus een overdrachtsfunctie met een helling van -6 dB/octaaf en




Figuur 3.6: *Schematisch overzicht van de toepassing van pre-emphase en de-emphase filtering bij de analyse respectievelijk de resynthese voor een stemhebbend spraakfragment.*

zorgt ervoor dat de gesynthetiseerde spraak weer een globale helling van -6 dB/octaaf vertoont. Het ($1^e$-orde) de-emphase filter en het ($10^e$-orde) synthesefilter worden nu gecombineerd tot één geheel, waarmee een $11^e$-orde filter ontstaat. Dit productiefilter is dan de geïnverteerde van het pre-emphase- en analysefilter tesamen, waarvan de filtercoëfficiënten $b_k$ gegeven worden door :

$$b_k = a_k + u a_{k-1} \qquad u = -0.9, \;\; a_{11} = 0, \;\; k = 1, \ldots, 11 \tag{3.20}$$

waarbij $a_k$ de filtercoëfficiënten van het analysefilter zijn. Het te synthetiseren signaal wordt zodoende i.p.v. door vgl. (3.9) gegeven door :

$$s'_n = G u_n - \sum_{k=1}^{11} b_k s'_{n-k} \tag{3.21}$$

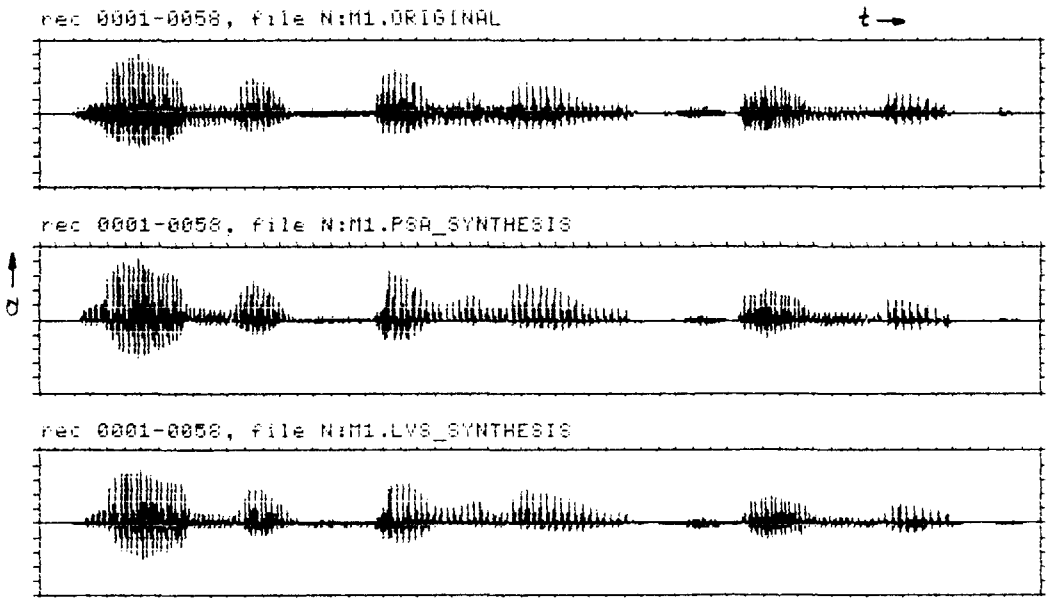
met de coëfficiënten $b_k$ gedefiniëerd volgens (3.20). De toepassing van pre-emphase en de-emphase filtering bij de analyse respectievelijk de resynthese is nog eens schematisch weergegeven in figuur 3.6 voor een stemhebbend spraakfragment.

Aldus worden voor iedere periode aan de hand van de ingelezen analyseparameters de samples van het te synthetiseren signaal stuk voor stuk volgens (3.21) berekend. De laatste 11 samples van de gesynthetiseerde periode worden daarbij telkens opgeslagen ten behoeve van de synthese van de volgende periode. De gesynthetiseerde samples worden per periode weggeschreven naar een nieuwe N-spraakfile. Vervolgens worden de parameters van de volgende periode ingelezen en wordt de bovenbeschreven synthese-cyclus herhaald.

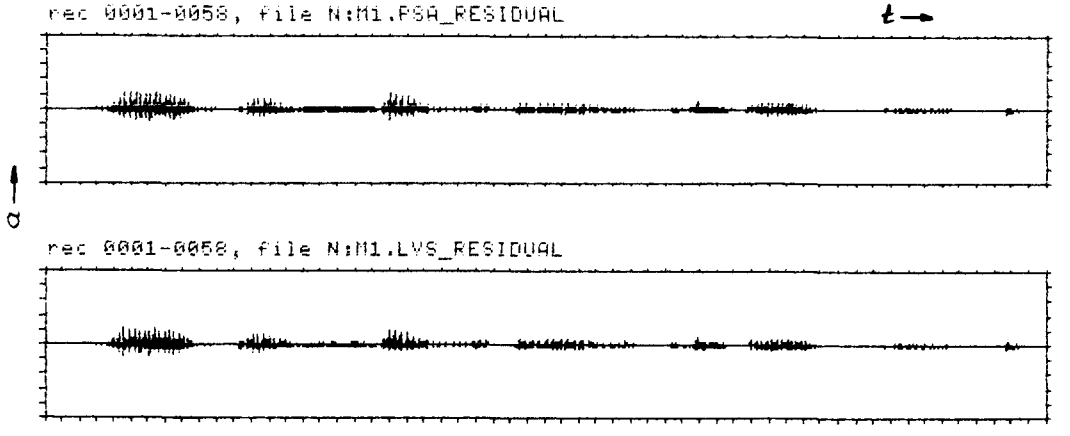
Nadat zo alle periodes gesynthetiseerd en weggeschreven zijn, wordt de nieuwe spraakfile nog genormeerd op de maximale absolute waarde van 2047 (12 bits), zodat de gecreëerde spraakfile compatibel is met het LVS-systeem. Ter evaluatie van het gesynthetiseerde spraakgeluid kan dit nu via het spraakuitgifte-systeem ten gehore worden gebracht of kan de golfvorm van het signaal (amplitude versus tijd) grafisch worden weergegeven. Dit laatste is gedaan in figuur 3.7, waar de golfvorm van de pitch-synchroon gesynthetiseerde spraakuiting *"Maandag gaan we naar het zwembad"*, uitgesproken door een mannenstem (spreker 1,zie appendix A), is weergegeven tesamen met de golfvorm van het originele spraakgeluid én de golfvorm van het met het LVS-systeem gesynthetiseerde spraaksignaal. In figuur 3.8 is bovendien het residusignaal van de pitch-synchrone analyse weergegeven naast het residusignaal van de LVS-analyse. Het eerstgenoemde signaal is hierbij eveneens pitch-synchroon berekend volgens vgl. (3.1) door telkens de responsie van het pitch-synchroon berekende analysefilter op één pitch-periode van het originele spraaksignaal te bepalen.

Zowel uit figuur 3.7 als uit figuur 3.8 blijkt, in deze vorm weergegeven, het ontwikkelde pitch-synchrone analyse-resynthese systeem vrijwel hetzelfde resultaat te geven als het conventionele LVS-systeem. Ook perceptief lijkt geen verschil hoorbaar te zijn tussen de met de respectievelijke systemen gesynthetiseerde spraakgeluiden. Een nadere perceptieve evaluatie zal echter in hoofdstuk 5 plaatsvinden.

Een overzicht van de software-programma's, die ontwikkeld zijn als basis van het in dit hoofdstuk beschreven pitch-synchrone analyse-resynthese systeem, is opgenomen in appendix D.




Figuur 3.7: *Golfvorm van de spraakuiting "Maandag gaan we naar het zwembad", uitgesproken door een mannenstem (spreker 1,zie App.A). Boven : originele spraak, midden : pitch-synchroon gesynthetiseerde spraak, onder : LVS-gesynthetiseerde spraak.*




Figuur 3.8: *Residusignaal na analyse van de originele spraakuiting van figuur 3.7. Boven : na pitch-synchrone analyse, onder : na LVS-analyse.*

# Hoofdstuk 4

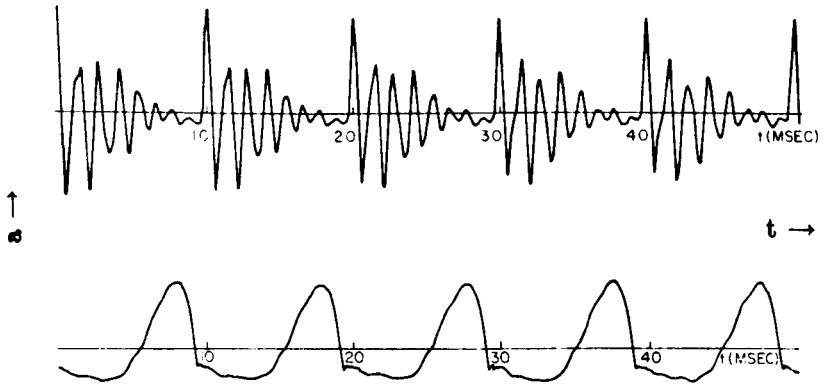
# Manipulatie van de excitatiefunktie

## 4.1 Inleiding

Zoals reeds in de inleiding van dit rapport is vermeld, was het doel van dit afstudeerproject te onderzoeken of de natuurlijkheid van LPC-gesyntheti seerde spraak verbeterd kan worden door de implementatie van een andere excitatiefunctie, waarmee het LPC-filter in het geval van stemhebbende spraakfragmenten geëxciteerd wordt. Het in het vorige hoofdstuk beschreven pitch-synchrone analyse-resynthese systeem geeft ons nu de mogelijkheid om, na analyse van een spraakuiting, het effect van het gebruik van verschillende excitatiefuncties bij de resynthese te onderzoeken. We kunnen met dit systeem immers de excitatiefunctie per periode wijzigen.

In het huidige LVS-systeem wordt als excitatiepuls voor stemhebbende gedeelten een eenheidspuls (deltapuls) met herhalingsfrequentie $F_0$ gebruikt. Conform de LPC-analyse-resynthese-techniek (§ 3.3) heeft deze funktie een vlak spectrum.

De wijziging van deze excitatiepuls bestaat nu hierin, dat een minder geïdealiseerde functie dan een eenheidspuls gebruikt wordt, die het werkelijke excitatiesignaal van de stembanden beter benadert. Het idee hierbij was om allereerst een functie met een bepaalde *vaste* vorm te gebruiken. Hiervoor worden in de literatuur verschillende parametrische modellen genoemd. Daarnaast was het de bedoeling een functie te implementeren, waarvan bepaalde parameters *variabel* in de tijd zijn, waarbij de waar-




Figuur 4.1: *Voorbeeld van het glottale bronsignaal (onder), verkregen via inverse-filtering van een stemhebbend spraakfragment (boven) (Rosenberg,1971).*

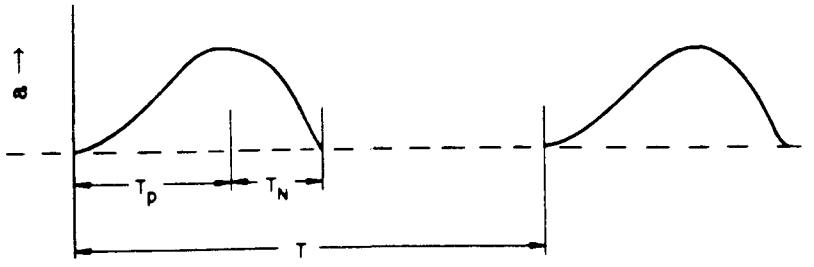
den van deze parameters uit het pitch-synchroon berekende residusignaal verkregen worden. Beide manipulaties zullen nu achtereenvolgens in dit hoofdstuk besproken worden.

De manipulaties betreffen alleen de *stemhebbende* gedeelten in de spraak. Het bronsignaal voor stemloze periodes blijft ongewijzigd ongecorreleerde witte ruis.

## 4.2 Excitatiefunctie beschreven door *vaste* parameters

### 4.2.1 Parametrisering van de glottale puls

Uit eerder onderzoek (Rosenberg,1971/Holmes,1973) is gebleken dat modellering van het bronsignaal, dat wordt opgewekt door de stembanden (glottale puls genoemd), belangrijk is voor het behoud van de natuurlijkheid in synthetische spraak. In de literatuur worden verschillende mathematische modellen voorgesteld, die de meest essentiële eigenschappen van de glottale puls beschrijven. Deze modellen zijn gebaseerd op onderzoek naar het glottale bronsignaal d.m.v. optische en electroglottografische methoden (Flanagan,1972), waarbij de stembandbewegingen gemeten worden, of op




Figuur 4.2: *Definitie van de glottale puls parameters in het 'Rosenberg-model' (Rosenberg,1971).*

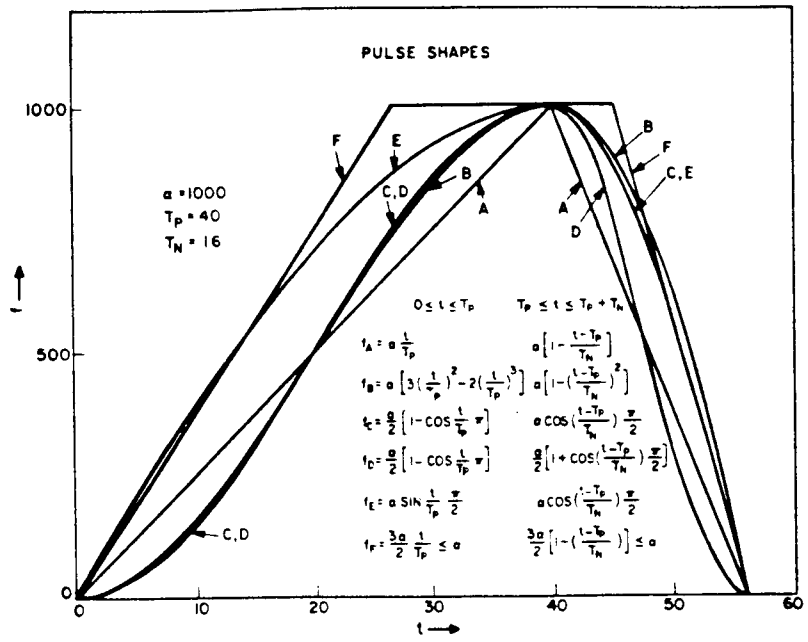
inverse filtering [1] van natuurlijke spraaksignalen (Holmes,1973). In figuur 4.1 is een voorbeeld van het glottale bronsignaal voor een stemhebbend spraakfragment weergegeven.

Zo wordt in het 'Rosenberg-model' (Rosenberg,1971) de glottale puls beschreven door twee parameters, zoals is weergegeven in figuur 4.2. De momentane pitch-periode is hierin aangegeven door $T$. De parameter $T_P$ (glottale openingstijd) is het gedeelte van de puls met een positieve helling en komt overeen met de tijd dat de stembanden zich openen. De glottale sluitingstijd $T_N$ is het gedeelte van de puls met een negatieve helling, waarbij de stembanden zich weer sluiten. De amplitude van de pulsen wordt constant gehouden en vormt dus geen experimentele parameter. De puls kan zodoende gespecificeerd worden door de relatieve openings- en sluitings tijden $T_P/T$ en $T_N/T$.

Rosenberg onderzocht nu zes verschillende pulsvormen (beschreven door polynomen of sinusoïden, zie figuur 4.3), die allen een sterke gelijkenis vertoonden met de natuurlijke glottale golfvorm. De pulsen hadden allen een relatieve openings- en sluitingstijd van 40% respectievelijk 16%, ze verschilden echter onderling in het aantal en de locatie van discontinuïteiten in hun helling (1e afgeleide). Perceptieve testen toonden aan dat pulsvormen met één discontinuïteit in de afgeleide op de plaats waar de stembanden zich gesloten hebben, het meest geprefereerd worden. Tussen de pulsen die hieraan voldoen bleek onderling weinig tot geen verschil hoorbaar te zijn (Sambur et al.,1978).

Andere modellen zijn in principe allen gebaseerd op het 'Rosenberg-

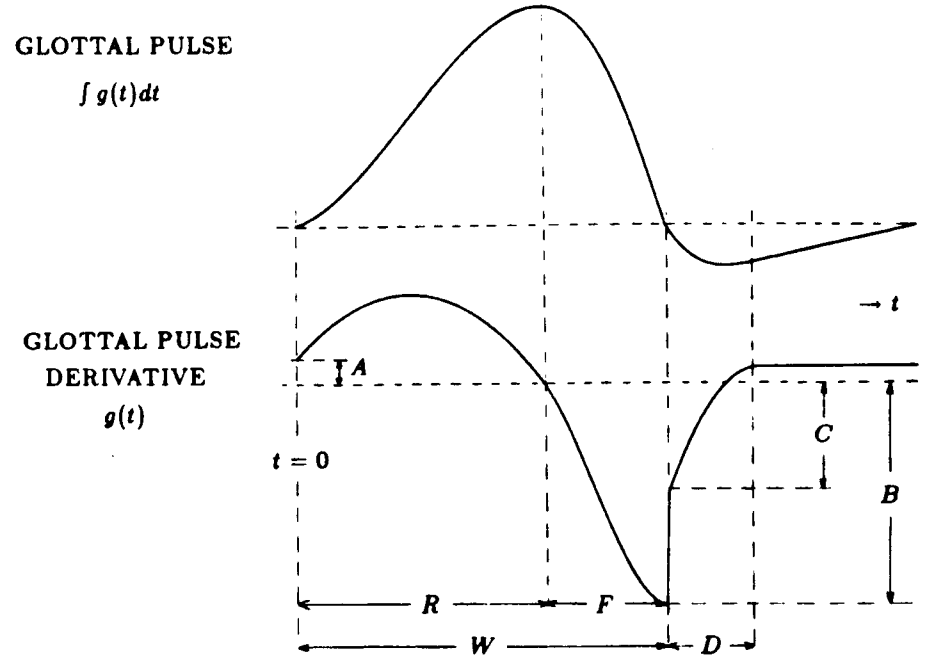
[1] De techniek van invers filteren zal in § 4.3 ter sprake komen.




Figuur 4.3: *Pulsvormen, zoals gebruikt door Rosenberg (Rosenberg,1971).*

model', maar voegen elk bepaalde verfijningen toe, zoals de mogelijkheid tot het onafhankelijk variëren van de plaats van de genoemde discontinuïteit, de introductie van een negatief puls-gedeelte na het sluiten van de stembanden of een afronding van de discontinuïteit.

Fujisaki & Ljungqvist (Fujisaki & Ljungqvist,1986 en 1987) zijn via een evaluatie en classificatie van de bestaande modellen tot een nieuw gegeneraliseerd model gekomen, dat alle essentiële aspecten van de afzonderlijke modellen omvat. In dit parametrisch model wordt de afgeleide van de glottale puls beschreven door opeenvolgende segmenten van polynomen, zoals is weergegeven in figuur 4.4. Het model kent drie tijdsparameters : de glottale openingstijd $R$, de glottale sluitingstijd $F$ (samen de open-pulsduur $W$) en het tijdsinterval tussen de glottale sluiting van de stembanden en de maximale negatieve waarde van de glottale puls ($D$). Daarnaast bevat het model nog drie amplitudeparameters die respectievelijk de helling van de puls bij het openen van de stembanden ($A$), de helling vóór het sluiten van de stembanden ($B$) en de helling direct ná het sluiten van de stembanden ($C$) bepalen. Het Fujisaki-Ljungqvist model biedt de mogelijkheid tot het onderling onafhankelijk variëren van de amplitude, de breedte en de helling van de puls alsmede de discontinuïteit in de afgeleide van de puls.




GLOTTAL PARAMETERS

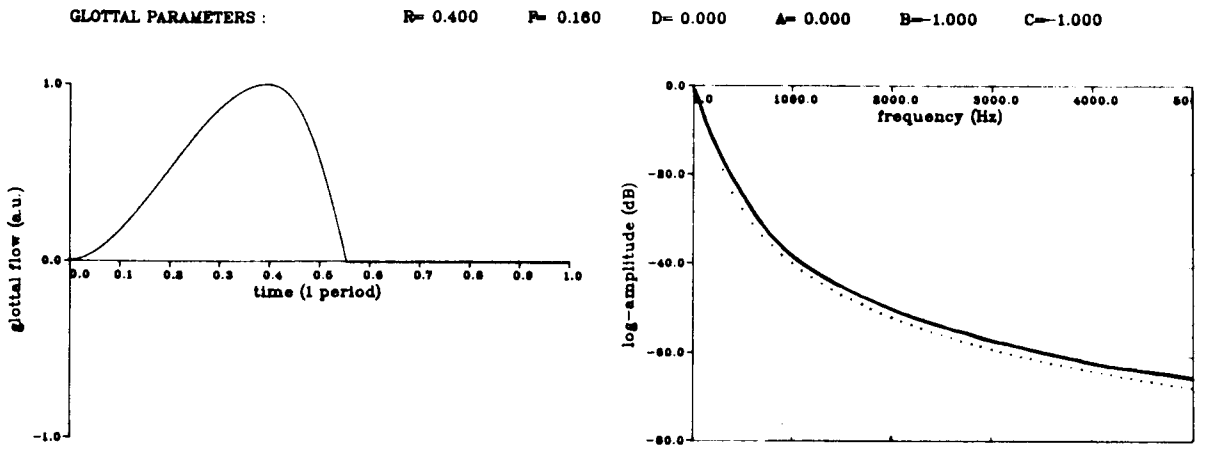
| | |
|---|---|
| W - PULSE WIDTH (R+F) | A - SLOPE AT GLOTTAL OPENING |
| S - PULSE SKEW (R+F)/(R-F) | B - SLOPE PRIOR TO CLOSURE |
| D - GLOTTAL CLOSURE TIMING | C - SLOPE FOLLOWING CLOSURE |

$$g(t) = \begin{cases} A - \frac{2A+R\alpha}{R}t + \frac{A+R\alpha}{R^2}t^2 & 0 < t \leq R \\ \alpha(t-R) + \frac{3B-2F\alpha}{F^2}(t-R)^2 - \frac{2B-F\alpha}{F^3}(t-R)^3 & R < t \leq W \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2 & W < t \leq W+D \\ \beta & W+D < t \leq T \end{cases}$$

waarin $\alpha := \frac{4AR+6FB}{2R^2-F^2}$ en $\beta := \frac{CD}{D-3(T-W)}$,

$T$ = grondtoonperiode.

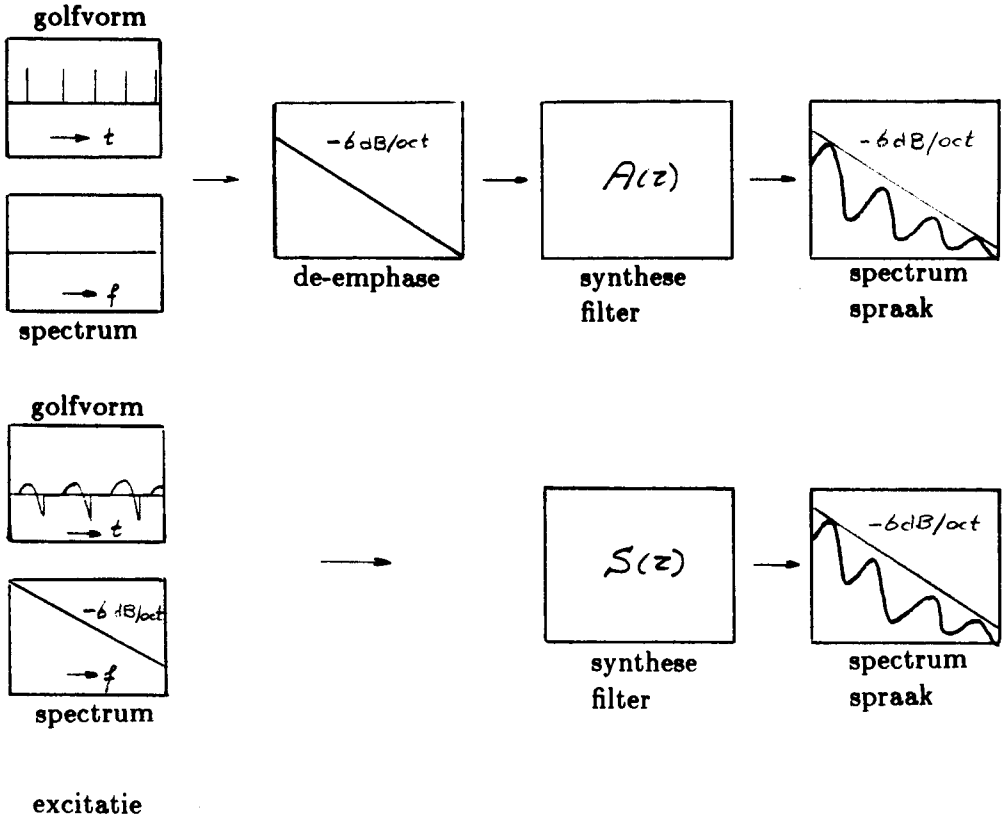
Figuur 4.4: *Definitie van golfvorm en parameters van de glottale puls volgens het Fujisaki-Ljungqvist model (Fujisaki & Ljungqvist,1986).*




Figuur 4.5: *Golfvorm (links) en omhullende van het amplitudespectrum (rechts) van de geparametriseerde glottale puls volgens de in de tekst genoemde keuze van de set glottale parameters. De gestippelde curve in het spectrum representeert een verloop van exáct -12 dB/octaaf.*

### 4.2.2 Implementatie in het resynthese-systeem

Uitgaande van het Fujisaki-Ljungqvist model kiezen we nu voor onze resynthese de volgende waarden voor de glottale parameters. In navolging van Rosenberg (Rosenberg,1971) wordt de openingstijd $R$ gelijk aan $0.4T$ respectievelijk de sluitingstijd $F$ gelijk aan $0.16T$ genomen (met $T$ de grondtoonperiode). De amplitude van de puls wordt constant gesteld, zodat de puls geschaald kan worden door $B = -1.0$ te kiezen. De parameters $A$, $C$ en $D$ worden zó gekozen, dat de verfijningen, die deze parameters beschrijven, (nog) niet meegenomen worden, d.w.z. $A = 0$, $D = 0$ en $C = B = -1.0$. Deze keuze van $C$ en $D$ zorgt er tevens voor, dat de zo geparametriseerde glottale puls de juiste spectrale omhullende heeft. Zoals in hoofdstuk 2 is beschreven, vertoont het natuurlijke bronsignaal immers een -12 dB/octaaf afval. De geparametriseerde glottale puls zal hier dus ook aan moeten voldoen. Nu geldt in het algemeen dat periodieke functies met één of meerdere discontinuïteiten in de 1e afgeleide (hieraan voldoen pulsvormen, gebaseerd op het Fujisaki-Ljungqvist model), een gemiddelde spectrale afval van -12 dB/octaaf vertonen. Door variaties van met name de parameters $C$ en $D$ kan de spectrale omhullende van de puls echter nog nader aangepast worden aan de gewenste -12 dB/octaaf afval. De genoemde keuze van $C$ en $D$ blijkt hierbij het beste resultaat te geven. Ter illustratie hiervan is in figuur 4.5 het amplitudespectrum weergegeven van de glottale puls volgens




Figuur 4.6: *Schematische weergave van de implementatie van de excitatiefunctie beschreven door vaste parameters (onder). Ter vergelijking is (boven) tevens het gebruik van de eenheidsimpuls als excitatiefunctie (zoals beschreven in hoofdstuk 3) geïllustreerd.*

de gekozen set glottale parameters.

In het bron-filter model voor spraakproduktie (zie hoofdstuk 2) onderscheiden we (voor stemhebbende spraakfragmenten) een bronsignaal met een spectrale omhullende van -12 dB/octaaf, een spraakproduktiefilter (met een vlakke omhullende) en een uitstralingseffect van +6 dB/octaaf. De bovenbeschreven parametrisering betreft de modellering van het bronsignaal. Willen we deze geparametriseerde glottale puls nu in ons analyse-resynthese systeem implementeren, dan moet het uitstralingseffect van +6 dB/octaaf nog in het synthese-proces verdisconteerd worden. Dit wordt bereikt door als excitatiefunctie voor het synthesefilter de afgeleide van de geparametriseerde glottale puls te gebruiken. Differentiëren in het tijddomein van

een digitaal opgeslagen signaal betekent immers een +6 dB/octaaf stijging in het frequentiedomein. Het uitstralingseffect wordt zodoende bij het bronsignaal betrokken. In figuur 4.6 is de implementatie schematisch weegegeven. Ter vergelijking is eveneens het in hoofdstuk 3 beschreven synthese-schema weergegeven, waarbij een eenheidsimpuls als excitatiefunctie gebruikt wordt.

Een tweede wijziging in het pitch-synchrone syntheseproces betreft de berekening van de amplitudeversterkingsfactor $G_v$ voor stemhebbende periodes. In § 3.5 is beschreven hoe $G$ berekend kan worden door de gemiddelde energie van het bronsignaal gelijk te stellen aan de gemiddelde energie van het residusignaal. Dit leidde voor stemloze periodes tot een amplitudefactor $G_u$, gegeven door vgl. (3.19) en voor stemhebbende periodes tot een amplitudefactor $G_v$ die, in het geval van een eenheidsimpuls als excitatiefunctie, gegeven wordt door vgl. (3.16). Willen we nu de boven beschreven geparametriseerde puls als excitatiefunctie gebruiken, dan wordt de gemiddelde energie van de stemhebbende excitatie echter niet langer gegeven door vgl. (3.15). Daarom wordt in dit geval een andere, tevens meer directe methode toegepast. Door namelijk tevens het residusignaal van de pitch-synchrone analyse te bepalen (zie § 3.5), kan de totale energie van het excitatiesignaal binnen de betreffende pitch-periode direct gelijkgesteld worden aan de totale energie van het residusignaal binnen dezelfde periode. Als excitatiesignaal $u_n$ moet hierbij, conform de LPC-analyse-resynthese techniek, de *tweemaal* gedifferentiëerde glottale puls (vlakke spectrale omhullende) genomen worden. Wordt het residusignaal verder gegeven door $e_n$, dan volgt zodoende de stemhebbende amplitudefactor $G_v$ uit :

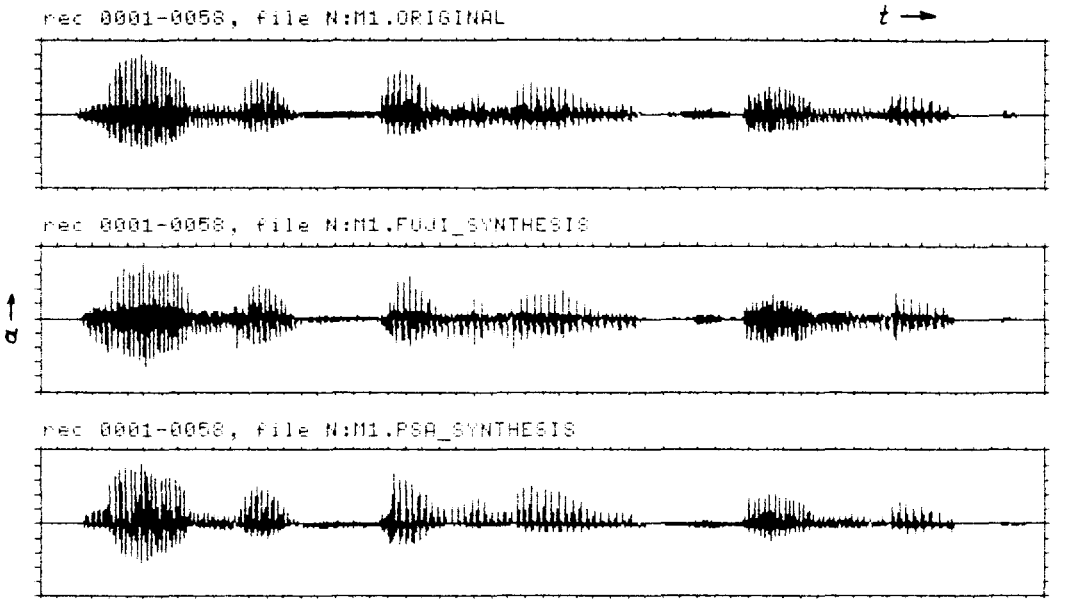
$$\sum_{n=n_0}^{n_1} e_n^2 = \sum_{n=n_0}^{n_1} (G_v u_n)^2 \tag{4.1}$$

oftewel :

$$G_v = \sqrt{\frac{\sum_{n=n_0}^{n_1} e_n^2}{\sum_{n=n_0}^{n_1} u_n^2}} \tag{4.2}$$

Hierbij zijn $n_0$ en $n_1$ het begin- resp. eindpunt (in samples) van de te synthetiseren periode.

De stemloze amplitudefactor $G_u$ wordt ongewijzigd bepaald door de methode van § 3.5 .




Figuur 4.7: *Golfvorm van de spraakuiting "Maandag gaan we naar het zwembad". Boven : originele spraaksignaal. Midden : pitch-synchroon gesynthetiseerd m.b.v. excitatiefunctie volgens Fujisaki-Ljungqvist model. Onder : pitch-synchroon gesynthetiseerd m.b.v. eenheidsimpuls als excitatiefunctie.*

Voor het overige wordt het pitch-synchrone syntheseproces uitgevoerd zoals beschreven in § 3.5 . In figuur 4.7 (midden) is de golfvorm weergegeven van een spraaksignaal, dat gesynthetiseerd is met de in deze paragraaf besproken excitatiefunctie. Het betreft dezelfde spreker en dezelfde spraakuiting als in figuur 3.7. Ter vergelijking is nogmaals de golfvorm weergegeven van het originele spraaksignaal en van het synthetische spraaksignaal, waarbij een eenheidsimpuls als excitatiefunctie is gebruikt. Het perceptieve effect van het gebruik van de in deze paragraaf besproken excitatiefunctie zal in hoofdstuk 5 besproken worden.

## 4.3 Excitatiefunctie beschreven door *variabele* parameters

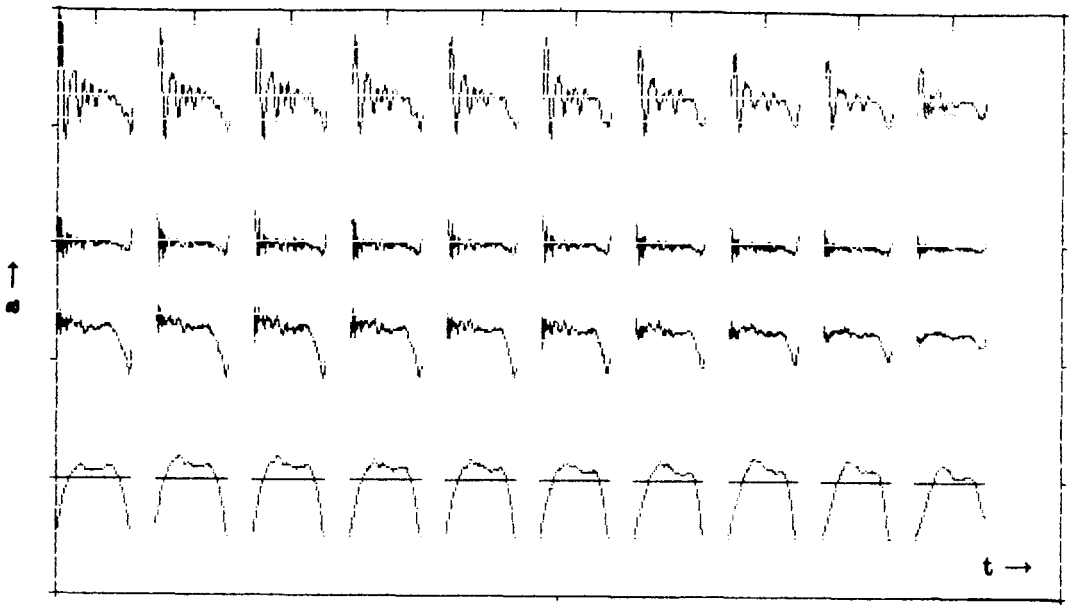
### 4.3.1 Inverse filtering

In de vorige paragraaf is de implementatie besproken van een excitatiefunctie, die gebaseerd was op een parametrisch model voor de glottale puls. Hierbij werd de excitatiefunctie vastgelegd door een *vaste* keuze van de parameters van het model. Daarnaast zouden we graag een excitatiefunctie toepassen, die beschreven wordt door parameters, die gebaseerd zijn op het natuurlijke spraaksignaal zelf, zodat ook de tijdsvariatie van de glottale puls in de modellering meegenomen kan worden.

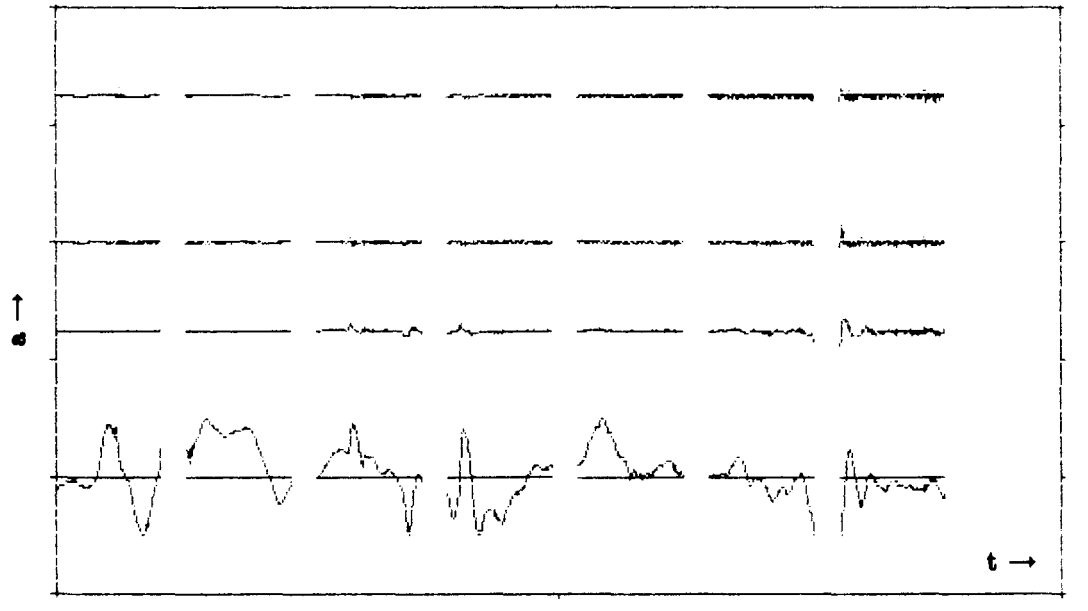
Informatie over het bronsignaal kan uit het natuurlijke spraaksignaal verkregen worden door de zgn. techniek van *invers filteren* (Markel & Gray,1976), die gebaseerd is op het bron-filter model voor spraakproduktie. Filteren we namelijk het natuurlijke spraaksignaal met de geïnverteerde van het spraakproduktiefilter, dan verkrijgen we als output een signaal, dat in principe de filter*excitatie* representeert.

In termen van ons pitch-synchrone LPC-analyse-resynthese systeem betekent dit dus het bepalen van het residusignaal door filtering van het natuurlijke spraaksignaal met het analysefilter. Het zo verkregen residusignaal representeert echter nog niet direct het bronsignaal, waarover we informatie willen verkrijgen. Allereerst moet het effect van de pre-emphase filtering, die bij de bepaling van het residusignaal aan het analysefilter voorafging, gecorrigeerd worden. Dit wordt bereikt door het residusignaal éénmaal te integreren, wat tot een -6 dB/octaaf verandering in het frequentiedomein leidt. Volgens het bron-filter model (zie hoofdstuk 2) is in het dan verkregen signaal ook nog het uitstralingseffect van +6 dB/octaaf verdisconteerd. De gezochte representatie van het bronsignaal wordt daarom uiteindelijk verkregen door het éénmaal geïntegreerde residusignaal nog éénmaal te integreren (-6 dB/octaaf verandering). Het zo verkregen tweemaal geïntegreerde residusignaal vertoont dan een spectrale omhullende van -12 dB/octaaf, wat overeenkomt met de spectrale helling van het bronsignaal.

Essentiëel bij de bovenbeschreven procedure is dat ons analyse-resynthese systeem *pitch-synchroon* werkt. Alleen indien het tweemaal geïntegreerde residusignaal per pitch-periode bepaald wordt, verkrijgen we een periodiek signaal, waarvan één periode geïdentificiëerd kan worden met de eerder




Figuur 4.8: *Voorbeeld van het pitch-synchrone inverse-filtering proces voor een stemhebbend spraakfragment. V.b.n.o. : het natuurlijke spraaksignaal, het residusignaal, het éénmaal geïntegreerde en tweemaal geïntegreerde residusignaal.*




Figuur 4.9: *Idem als figuur 4.8, nu echter voor een stemloos spraakfragment.*

genoemde glottale puls. In dat geval levert het tweemaal geïntegreerde residusignaal de gezochte informatie, waarop we de te gebruiken excitatiefunctie kunnen baseren.

In figuur 4.8 is een voorbeeld gegeven van het resultaat van bovengenoemde procedure. Het betreft een stemhebbbend spraakfragment ter lengte van enkele pitch-periodes. Bovenin de figuur is per periode de golfvorm van het originele, natuurlijke spraaksignaal weergegeven (hierbij wordt overigens nogmaals het uitsnijden van de pitch-periodes geïllustreerd). Van boven naar beneden zijn vervolgens per periode weergegeven : het pitch-synchroon berekende residusignaal $e_n$, het éénmaal geïntegreerde residusignaal $e'_n$ en tenslotte het tweemaal geïntegreerde residusignaal $e''_n$. Het laatste signaal is hierbij per periode geschaald. Het betreft in deze figuur allemaal *stemhebbende* pitch-periodes. Het tweemaal geïntegreerde residusignaal $e''_n$ blijkt inderdaad een duidelijke periodieke pulsstructuur te vertonen. Ter vergelijking zijn in figuur 4.9 dezelfde signalen weergegeven, maar nu voor een *stemloos* spraakfragment. Hierin is te zien dat nu, zoals verwacht, een periodieke pulsstructuur duidelijk ontbreekt.

Vergelijken we het tweemaal geïntegreerde residusignaal, zoals dat van figuur 4.8, met de glottale pulsvorm volgens het in § 4.2.1 beschreven (Fujisaki-Ljungqvist) model, dan zijn er duidelijk twee verschillen te onderscheiden.

Allereerst vertoont het tweemaal geïntegreerde residusignaal eerder overeenkomst met de afgeleide van de Fujisaki-Ljungqvist puls dan met de glottale puls zelf. Een mogelijke verklaring hiervan zou gegeven kunnen worden door het feit dat bij de opname van het originele spraaksignaal de laagste frequenties van het door de spreker/spreekster uitgesproken spraakgeluid niet in de opname worden meegenomen. Een dergelijke high-pass-filtering resulteert immers in een 'differentiërende' werking.

Een tweede verschil betreft de locatie van de puls binnen een pitch-periode. De in § 4.2 besproken excitatiefunctie werd zodanig binnen een pitch-periode geplaatst, dat het begin van de pitch-periode overeenkwam met het openen van de stembanden (zie bv. figuur 4.2 en 4.5). Het punt waar de stembanden zich sluiten (=plaats van de discontinuïteit in de afgeleide van de glottale puls) bevond zich zodoende ongeveer in het midden van de betreffende periode. In figuur 4.8 zien we echter dat dit punt nu juist aan het begin van de pitch-periode ligt. We kunnen nu dan ook een duidelijke interpretatie geven aan de eerder genoemde discontinuïteit in de

afgeleide van de glottale puls, overeenkomend met het sluitingspunt van de stembanden in de glottale puls. Zoals ook in figuur 4.8 is te zien, veroorzaakt de discontinuïteit in de afgeleide van de glottale puls (in figuur 4.8 : het éénmaal geïntegreerde residusignaal) nu juist de primaire excitatie in de golfvorm van het spraaksignaal. De manier, waarop de in § 4.2 beschreven excitatiefunctie binnen een pitch-periode geplaatst werd, leidt er dus toe dat de excitatie in de golfvorm van het gesynthetiseerde spraaksignaal zich telkens ongeveer in het midden van de betreffende pitch-periode bevindt. Het gesynthetiseerde spraaksignaal vertoont zodoende een kleine verschuiving ten opzichte van het originele spraaksignaal.

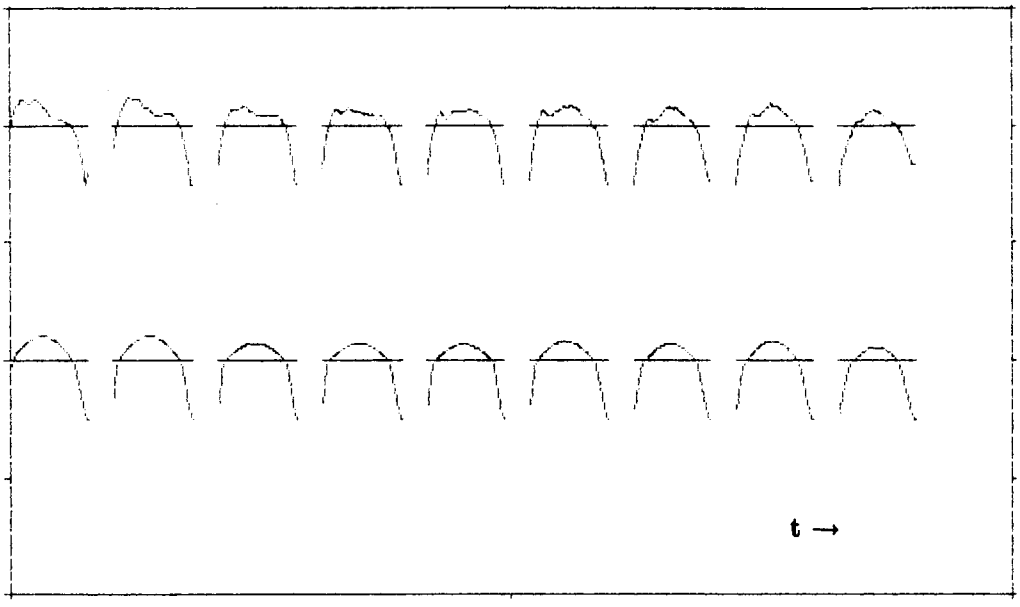
Tenslotte zien we in figuur 4.8 dat het tweemaal geïntegreerde residusignaal nog een zekere fijnstructuur vertoont. Dit is ook te begrijpen indien we beseffen dat het tweemaal geïntegreerde residusignaal slechts *modelmatig* (namelijk volgens het bron-filter model) het bronsignaal representeert. In werkelijkheid uiten eveneens alle afwijkingen van dit model (met name bron-filter interacties) zich in dit signaal. Bovendien kan het signaal ook nog enige filterkarakteristiek bevatten, voorzover een $10^e$-orde analysefilter niet toereikend was om deze te compenseren.

### 4.3.2 Stilering

Het tweemaal geïntegreerde residusignaal vertoont, in tegenstelling tot het éénmaal geïntegreerde residusignaal en het residusignaal zelf, een vrij regelmatig patroon. Het ligt daarom voor de hand de gezochte variabele excitatiefunctie te baseren op een stilering van het tweemaal geïntegreerde residusignaal m.b.v. een geschikte stileringsfunctie. De waarden van de parameters die de stileringsfunctie vastleggen worden dan voor iedere periode uit het tweemaal geïntegreerde residusignaal bepaald. Zodoende ontstaat een parametrisering die eveneens de tijdsontwikkeling van het bronsignaal omvat.

Volledig analoog aan de implementatie van de geparametriseerde glottale puls, zoals beschreven in § 4.2, kan het gestileerde tweemaal geïntegreerde residusignaal door tweemaal differentiatie weer getransformeerd worden tot een voor ons analyse-resynthese systeem geschikte excitatiefunctie. Ook de in § 4.2.2 beschreven methode ter berekening van de amplitudefactor $G_v$ voor stemhebbende periodes wordt hier toegepast.

Het probleem bij het ontwikkelen van een variabele excitatiefunctie




Figuur 4.10: *Stilering van het tweemaal geïntegreerde residu van een stemhebbend spraakfragment m.b.v. een sinusfunctie en rechten. Boven : tweemaal geïntegreerde residusignaal $e''_n$. Onder : gestileerde signaal $f_n$.*

spitst zich dus toe op het vinden van een juiste stileringsfunctie(s). Hiertoe zijn in dit onderzoek verscheidende mogelijkheden onderzocht.

Een mogelijke stilering is weergegeven in figuur 4.10. Boven in deze figuur is het tweemaal geïntegreerde residusignaal $e''_n$ weergegeven (wederom per periode geschaald), zoals dat via de besproken inverse-filtering procedure uit een stemhebbend spraakfragment is verkregen. Onder in de figuur is nu (eveneens geschaald) het gestileerde signaal $f_n$ weergegeven. De stileringsfunctie bestaat hier uit een halve sinusperiode tussen de nulpunten van het signaal $e''_n$ en uit rechten buiten deze nulpunten. De stileringsfunctie wordt beschreven door 7 parameters, te weten : de plaats van de nulpunten, de amplitude (hoogte) van de sinusfunctie en de plaats en waarden (diepte) van de twee minima. De plaatsen van de minima worden ook vastgelegd omdat deze niet altijd samen blijken te vallen met het begin- resp. eindpunt van de betreffende periode. Implementatie van het zo gestileerde residusignaal in het resynthese-systeem blijkt echter in een slechte spraakkwaliteit te resulteren. Nader onderzoek naar de oorzaak hiervan wees uit dat aan de te gebruiken stileringsfunctie de eis gesteld moet worden, dat zijn afgeleide binnen de pitch-periode geen discontinuïteiten

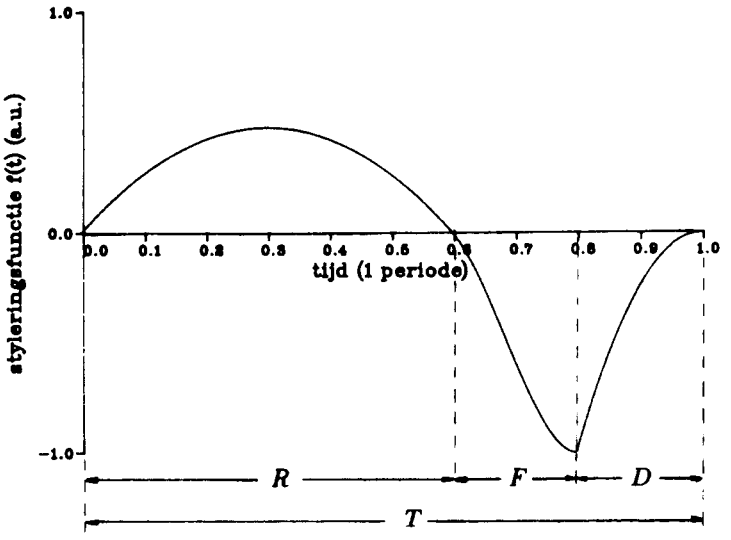
bevat. We hebben immers reeds eerder gezien, dat een discontinuïteit in de afgeleide van het bronsignaal tot een excitatie in de golfvorm van het uiteindelijke spraaksignaal leidt. Bekijken we nu de stileringsfunctie, zoals weergegeven in figuur 4.10, dan zien we dat dergelijke discontinuïteiten, behalve op de overgang tussen twee periodes, ook telkens ter plaatse van de twee nulpunten optreden. Dit leidt zodoende tot ongewenste secundaire excitaties in de golfvorm van het gesynthetiseerde signaal naast de primaire excitaties aan het begin van iedere periode.

De te gebruiken stileringsfunctie moet dus binnen een periode een continue afgeleide hebben. Een analoog resultaat werd gevonden door Rosenberg(1971), die vond dat glottale pulsvormen met één discontinuïteit het meest geprefereerd worden (zie § 4.2.1) Daarnaast gelden een aantal randvoorwaarden, zoals de plaats van de twee nulpunten, de energie van het signaal tussen deze nulpunten en de waarden van het signaal in het begin- en eindpunt van de periode. Bovendien vertoont het signaal $e_n''$ een zekere asymmetrie (langzaam stijgende opgaande flank, steil dalende flank), die we ook graag in de stileringsfunctie zouden willen verdisconteren. Deze veelheid van eisen leidt tot de noodzaak van het toepassen van een samengestelde functie, beschreven door een opeenvolging van b.v. polynomen.

Enkele van dergelijke stileringsfuncties, die nog door een acceptabel aantal parameters beschreven kunnen worden, zijn onderzocht. Zo is een functie opgebouwd uit vier parabolen gebruikt, alsmede een functie gebaseerd op het Fujisaki-Ljungqvist model (zie figuur 4.4). Toch bleken ook deze stileringen geen goede synthetische spraakkwaliteit op te leveren.

Een mogelijke verklaring hiervoor zou kunnen liggen in de aansluiting van het gestileerde signaal op de overgangen tussen twee periodes. Hoewel het gestileerde signaal zelf wel aansluit op deze overgangen, kunnen er bij de synthese echter discontinuïteiten ontstaan, doordat het signaal voor iedere periode met een andere amplitudefactor $G$ vermenigvuldigd wordt. Bij de in § 4.2 beschreven 'vaste' excitatiefunctie was dit niet het geval, aangezien deze zodanig binnen een pitch-periode geplaatst werd, dat de functiewaarden in het begin- en eindpunt van de periode nul zijn. In dat geval is ook na vermenigvuldiging met $G$ de continuïteit gewaarborgd.

Daarom is uiteindelijk een stilering geïmplementeerd, waarbij het 1e nulpunt van het gestileerde signaal verschoven wordt naar het beginpunt van de betreffende periode. Voor de stileringsfunctie $f_n$ wordt hierbij een gewijzigde vorm van het functievoorschrift voor de afgeleide van de glot-



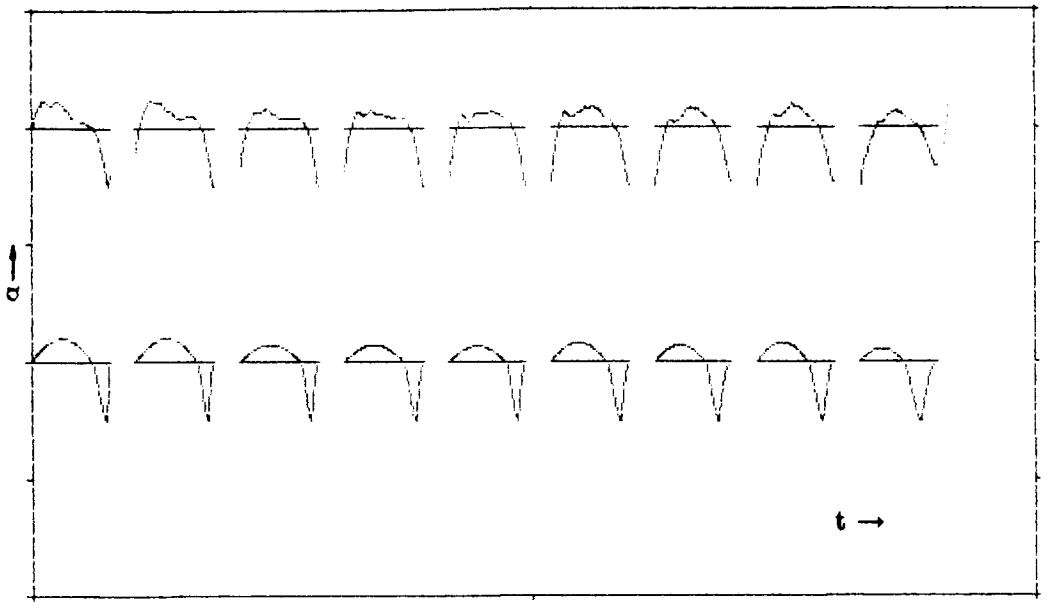

Figuur 4.11: *Voorbeeld van de gebruikte stileringsfunctie voor* $R = 0.6$, $F = 0.2$ *en* $H = 2.0$.

tale puls in het Fujisaki-Ljungqvist model gebruikt (functie $g(t)$ in figuur 4.4). Teneinde de functie zo goed mogelijk overeen te laten komen met het residusignaal $e_n''$, wordt allereerst $A := 0$ ($g(0) := 0$), $C := B = -1$ en $D := T - (R + F)$ genomen. Om nu de functiewaarde in het eindpunt van de periode eveneens nul te krijgen ($g(T) := 0$), wordt $\beta$ gelijk aan nul gesteld. Verder wordt een mogelijkheid geïntroduceerd voor het onafhankelijk variëren van de amplitude van de functie tussen de nulpunten, door in de parameter $\alpha$ een ampltudeparameter $H$ op te nemen volgens :

$$\alpha := H\frac{6FB}{2R^2 - F^2} \tag{4.3}$$

M.b.v. het functievoorschrift voor $g(t)$ (zie figuur 4.4) kan aangetoond worden dat deze introductie de continuïteit van de functie in $t = R$ en $t = R+F$, alsmede de continuïteit van de helling van de functie in $t = R$ niet aantast. Een mogelijk verloop van de aldus gedefiniëerde stileringsfunctie is weergegeven in figuur 4.11 (voor $R = 0.6$, $F = 0.2$ en $H = 2.0$).

De stileringsfunctie wordt zodoende beschreven door 3 variabele parameters : $R$, $F$ en $H$, waarvan de waarden per periode uit het residusignaal $e_n''$ bepaald worden. De parameter $R$ volgt uit de afstand tussen de twee nulpunten in het signaal $e_n''$, de parameter $F$ uit de afstand tussen de plaats van het 2e minimum (meestal het eindpunt van de periode) en het 2e nulpunt. De parameter $H$ wordt bepaald door de energie tussen de nulpunten in het signaal $e_n''$ gelijk te stellen aan de bijbehorende energie in het gestileerde signaal.

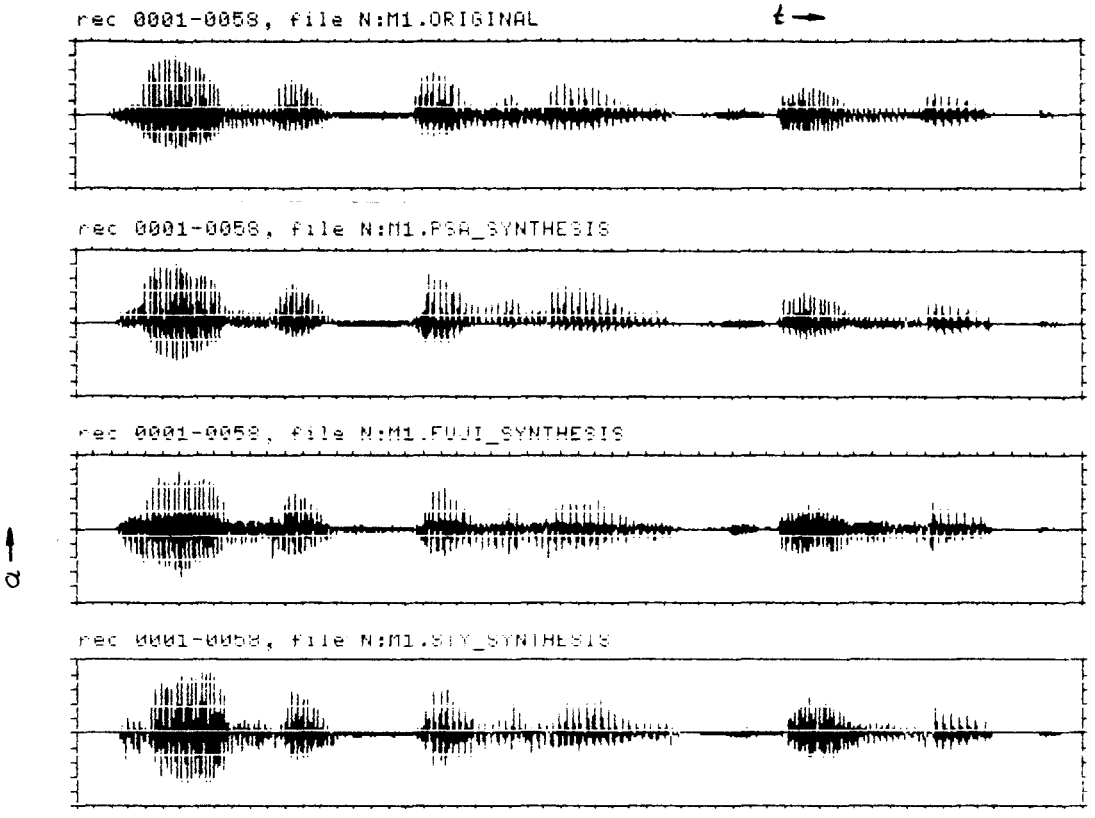



Figuur 4.12: *Gebruikte stilering van het tweemaal geïntegreerde residu van een stemhebbend spraakfragment m.b.v. de functie van figuur 4.11. Boven : tweemaal geïntegreerde residusignaal $e''_n$. Onder : gestileerde signaal $f_n$.*

Het resultaat van een dergelijke stileringsprocedure voor een stemhebbend spraakfragment is weergegeven in figuur 4.12. Bij de uitvoer van het stileringsproces op langere spraakuitingen bleek, met name op overgangen van stemhebbende naar stemloze gedeelten, de structuur van het tweemaal geïntegreerde residusignaal incidenteel minder duidelijk aanwezig. In de gevallen dat dit problemen gaf voor de bepaling van de parameters $R$, $F$ en $H$, is dat opgelost door de betreffende periode stemloos te maken (voor stemloze periodes vindt immers in het geheel geen stilering plaats).

Zoals door bestudering van figuur 4.12 is in te zien, moeten we wel beseffen dat de gebruikte stileringsmetode de relatieve oriëntatie van de excitatiepunten (plaats van de minima in het residusignaal $e''_n$) enigszins wijzigt. De stileringsfunctie wordt namelijk over de afstand tussen het beginpunt van een periode en het 1e nulpunt van het residusignaal $e''_n$, naar links verschoven. Deze afstand variëert echter van periode tot periode.

De kwaliteit van de, met het bovenbeschreven gestileerde residusignaal, gesynthetiseerde spraak laat echter toch nog te wensen over. Gedeeltelijk wordt dit waarschijnlijk veroorzaakt door de genoemde wijziging in de posities van de excitatiepunten. Daarnaast moeten we concluderen dat stileringen, zoals beschreven in deze paragraaf, blijkbaar toch niet voldoende




Figuur 4.13: *Golfvorm van de spraakuiting "Maandag gaan we naar het zwembad". Van boven naar onder : originele spraak ; pitch-synchrone synthese met eenheidsimpuls als excitatiefunctie ; idem met 'vaste excitatiefunctie ; idem met 'variabele' excitatiefunctie.*

essentiële eigenschappen van het residusignaal bevatten om spraak van de gewenste goede kwaliteit te verkrijgen. Toch is voor de bovenbesproken stileringsmethode gekozen, aangezien deze een relatief eenvoudige modellering biedt van de tijdsontwikkeling van het bronsignaal d.m.v. slechts drie parameters en voor korte spraakfragmenten wel goede resultaten geeft.

In figuur 4.13 (onder) is de golfvorm weergegeven van een spraaksignaal, dat gesynthetiseerd is aan de hand van het, op de bovenbeschreven manier gestileerde residusignaal. Het betreft weer dezelfde spreker en spraakuiting als van figuur 3.7 en figuur 4.7.

# Hoofdstuk 5

# Perceptieve evaluatie

## 5.1 Inleiding

Uitgaande van het pitch-synchrone analyse-resynthese systeem beschikken we nu, naast de conventionele IPO-LVS-spraak, over drie 'nieuwe' soorten synthetische spraak, die zich onderling onderscheiden in de gebruikte excitatiefunctie. De 'standaard' pitch-synchrone versie, zoals beschreven in hoofdstuk 3 (in het vervolg versie *PSA-standaard* genoemd), is overeenkomstig het LVS-systeem gebaseerd op een eenheidsimpuls als excitatiefunctie. Bij de twee in hoofdstuk 4 beschreven versies wordt een minder geïdealiseerde excitatiefunctie gebruikt in de vorm van een 'vaste' respectievelijk 'variabele' excitatiepuls (in het vervolg versie *PSA-vaste puls* respectievelijk versie *PSA-variabele puls* genoemd).

In dit hoofdstuk zal nu het perceptieve effect van het gebruik van deze verschillende excitatiefuncties (met name de mate van natuurlijkheid van de resulterende synthetische spraakversies) besproken worden.

## 5.2 Perceptief experiment

Bij het in eerste instantie informeel beluisteren van de drie versies pitch-synchrone spraak, ontstaat de indruk dat de standaard PSA-versie zich perceptief niet onderscheidt van de LVS-gesynthetiseerde spraak. De versie PSA-vaste puls blijkt van een goede spraakkwaliteit te zijn en lijkt in vergelijking met de standaard PSA-versie iets voller (donkerder) te klinken.

De versie PSA-variabele puls blijkt, zoals reeds gezegd in hoofdstuk 4, van een wat mindere kwaliteit dan de standaard PSA-versie.

Om echter wat meer kwantitatieve uitspraken te kunnen doen, zouden we een perceptief experiment willen uitvoeren, waarbij proefpersonen gevraagd wordt de diverse aangeboden synthetische spraakversies op hun natuurlijkheid te beoordelen. In de literatuur worden verschillende subjectieve preferentietests beschreven (voor een overzicht zie bv. Hovelynck,1985), waarbij het subjectieve waarde-oordeel van luisteraars gekwantificiëerd wordt d.m.v. schaalwaarden gelegen op een psychologisch continuüm.

Een probleem bij het testen van natuurlijkheid is echter de niet-eenduidigheid van het criterium 'natuurlijkheid'. Bij de interpretatie van de uit een test resulterende schaalwaarden moet men er daarom steeds op bedacht zijn of deze schaalwaarden wel de verschillende spraakversies representeren met betrekking tot het te onderzoeken criterium 'natuurlijkheid'.

Als experiment is gekozen voor de methode van paarsgewijze vergelijking, waarbij de te testen stimuli telkens in paren aan de proefpersonen worden aangeboden. De proefpersonen wordt gevraagd per paarvergelijking aan te geven welke van de twee aangeboden stimuli het meest natuurlijk overkomt. Doordat de proefpersonen steeds een duidelijke referentie hebben, kunnen met deze methode kleine verschillen tussen stimuli gemeten worden. Onlangs is op het IPO een model uit de literatuur (Scheffé,1952) uitgewerkt voor de analyse van data, afkomstig van een paarsgewijs vergelijkings-experiment (Damen & Ellermann,1988 / Damen, 1988). Dit model zal nu zeer in het kort beschreven worden. Voor een meer uitgebreide, algemene behandeling wordt verwezen naar Damen & Ellermann (1988) alsmede Damen (1988).

### 5.2.1 Een variantie-analyse voor paarsgewijze vergelijkingen

#### Het experiment

Uitgangspunt zijn $m$ te rangschikken items, die in $m(m-1)$ paren aan de proefpersonen worden aangeboden. De proefpersonen drukken hun voorkeur uit op een 3- of meerpuntsschaal. In onze experimenten is een 5-punts-scoreschaal gebruikt, waarbij de proefpersoon bij aanbieding van

het paar (i,j) kan kiezen uit :

(-2) *duidelijke voorkeur voor item i boven item j*

(-1) *lichtelijke voorkeur voor item i boven item j*

(0) *geen voorkeur*

(1) *lichtelijke voorkeur voor item j boven item i*

(2) *duidelijke voorkeur voor item j boven item i*

### Het wiskundig model

Het model van Scheffé is gebaseerd op een variantie-analyse van de resultaten van het experiment, waarbij statistische hypothesen getoetst worden. Het model kent een aantal belangrijke aannamen :

- Allereerst wordt verondersteld dat de voorkeursoordelen onafhankelijke variabelen zijn. Ze mogen alleen bepaald worden door het te onderzoeken effect (in ons geval : natuurlijkheid). Met name mag het beslist niet voorkomen dat proefpersonen de stimuli kunnen identificeren.

- Er is voldaan aan de zgn. *hypothese van subtractiviteit.* Dat wil zeggen dat het te onderzoeken effect op een ééndimensionaal psychologisch continuüm af te beelden moet zijn. Slechts in dat geval geldt dat de gemiddelde voorkeur van item i boven item j een maat is voor het verschil tussen de schaalwaarden behorend bij de betreffende items. Zodoende kunnen de schaalwaarden dan geschat worden uit de gemiddelde voorkeursscores.

- Er doen zich geen sterke *volgorde(orde-)effecten* voor, d.w.z. de gemiddelde voorkeur van item i boven item j wijkt niet al te veel af van het tegengestelde van de gemiddelde voorkeur van item j boven item i. De significantie van dergelijke orde-effecten moet dus getoetst worden.

- De voorkeursoordelen hebben dezelfde variantie (*homogeniteitshypothese*). Ook deze voorwaarde moet getest worden.

Is aan bovengenoemde voorwaarden voldaan, dan is het mogelijk de hoofdeffecten (de geschatte schaalwaarden) onderling te vergelijken en een uitspraak te doen of deze effecten al of niet significant zijn. De variantie van de geschatte schaalwaarden wordt hiertoe uitgedrukt in een zgn. 'yardstick', die het betrouwbaarheidsinterval aangeeft van het verschil van twee geschatte schaalwaarden. Is het verschil tussen twee schaalwaarden groter dan deze 'yardstick', dan mogen we concluderen dat de betreffende schaalwaarden significant verschillen.

### 5.2.2 Experiment 1

Met behulp van het model van Scheffé kunnen we nu een paarsgewijs vergelijkingsexperiment uitvoeren, waarbij het mogelijk zou moeten zijn om uitspraken te doen over de verschillen in natuurlijkheid van verschillende aangeboden items. Als items nemen we hierbij onze drie versies pitch-synchroon gesynthetiseerde spraak (versies : PSA-standaard, PSA-vaste puls, PSA-variabele puls).

#### Stimuli

Als spraakmateriaal zijn een aantal zinnen opgenomen, uitgesproken door 4 mannelijke en 4 vrouwelijke sprekers. De opnamen zijn gemaakt in de IPO-studio m.b.v. een Sony PCM-501 ES digitale audio recorder en een Brüel & Kjær condensator microfoon (type 4003). De opnamen zijn boven de 5 kHz low-pass gefilterd en met een bemonsteringsfrequentie van 10 kHz opgeslagen in de IPO-VAX/8530 computer m.b.v. een 16-bits analoog-digitaal converter .

Ten behoeve van het experiment is een stimuluszin uitgekozen, die relatief weinig stemloze spraakgedeelten bevat, namelijk de zin :

*"Maandag gaan we naar het zwembad."*

Centraal in dit onderzoek staat immers de excitatiefunctie voor *stemhebbende* gedeelten. Het stimulusmateriaal bestond zodoende uit één zin uitgesproken door 8 verschillende sprekers. In appendix A is een lijst van de gebruikte sprekers opgenomen. Tevens is hierin de golfvorm en het verloop van de grondtoon voor de diverse sprekers weergegeven. De zinnen zijn voor iedere spreker pitch-synchroon geanalyseerd (zoals beschreven in hoofdstuk 3) en

vervolgens zijn de drie versies pitch-synchroon gesynthetiseerde spraak volgens de in hoofdstuk 3 en 4 beschreven methoden gegenereerd. Tenslotte is het spraakniveau van deze versies gelijkgemaakt (volgens de EPL-methode van Brady,1968). Op deze manier zijn dus van iedere spreker 3 spraakstimuli beschikbaar :

- versie 1 : PSA-standaard
- versie 2 : PSA-vaste puls
- versie 3 : PSA-variabele puls

### Opzet van het experiment

Het experiment bestaat nu uit 8 deelexperimenten, waarbij telkens van één spreker de 6 mogelijke paarsgewijze combinaties van de 3 versies spraak aangeboden worden.

Om de proefpersonen een indruk te geven van de te beoordelen stimuli, worden telkens aan het begin van een deelexperiment (andere spreker) alle drie de versies éénmaal als proefstimuli aangeboden. Hierbij hoeven de proefpersonen geen responsie te geven.

Vervolgens worden de 6 paren stimuli voor de betreffende spreker aangeboden. De volgorde waarin dit gebeurt wordt zodanig gekozen, dat er zoveel mogelijk willekeur ontstaat en er zo weinig mogelijk identificatie mogelijk is (zie Damen,1988). Ook de volgorde van de deelexperimenten (sprekers) wordt zoveel mogelijk willekeurig gehouden.

Via het IPO-LVS-spraakuitgifte systeem worden de stimuli in de aldus verkregen volgorde m.b.v. een PCM-recorder op een geluidsband opgenomen. Bij de uitvoer van het experiment wordt deze band via koptelefoons (Pioneer,type Monitor 10) in een geluidsarme luisterruimte aan de proefpersonen ten gehore gebracht. Het experiment bestaat uit één sessie van $\pm$ 10 minuten.

### De proefpersonen

De proefpersonen die bij het experiment gebruikt zijn, zijn in twee groepen te onderscheiden :

- proefpersonen, die in meer of mindere mate gewend zijn aan kunstmatige spraak en geacht worden hier vrij kritisch naar te luisteren. Hiertoe behoort een groot deel van de medewerkers binnen de Akoestisch-Fonetische Groep.
- proefpersonen, die niet eerder of nauwelijks kunstmatige spraak gehoord hebben. Deze groep bestond voor het grootste gedeelte uit overige IPO-medewerkers.

In totaal namen er 20 personen deel aan het experiment, waarvan er 10 tot de geoefende en 10 tot de ongeoefende groep gerekend kunnen worden. Bij de verwerking van de data zijn de resultaten, behalve voor de totale groep proefpersonen, eveneens voor deze twee groepen afzonderlijk bepaald.

Verder is door geen van de proefpersonen een afwijking van het gehoor gerapporteerd.

## De instructie

De proefpersonen moesten voor ieder aangeboden paar op een antwoordformulier hun voorkeur voor één van beide stimuli weergegeven. Aan het einde van het experiment hadden de proefpersonen de gelegenheid om zowel mondeling als schriftelijk hun commentaar te geven. Een voorbeeld van een antwoordformulier is opgenomen in appendix B. Vóór de start van het experiment werd de proefpersonen via een instructieformulier de opzet van het experiment duidelijk gemaakt, waarbij geïnstrueerd werd de stimuli te beoordelen op basis van een algemene indruk van de natuurlijkheid. Het gebruikte instructieformulier is weergegeven in appendix C.

## Verwerking van de data

De responsies van de proefpersonen (in de vorm van voorkeursoordelen voor ieder paar (i,j) worden geordend tot zgn. frequentiematrices, waarin voor ieder paar (i,j) het aantal responsies per voorkeursoordeel is aangegeven. Voor iedere spreker worden drie van dergelijke frequentiematrices opgesteld : één gebaseerd op de responsies van *alle* proefpersonen en twee voor alleen de *geoefende* repectievelijk de *ongeoefende* proefpersonen.

In deze matrix-vorm zijn de data geschikt als input voor een door Damen en Ellermann (1988) ontwikkeld programma, dat als output de schaalwaar-

den behorende bij de drie geteste versies, alsmede een uitsluitsel over de significantie van de hoofdeffecten, mogelijke orde-effecten en afwijkingen van de subtractiviteits- en homogeniteitshypothese geeft.

**Resultaten**

Bij de verwerking van de responsies bleek er veelvuldig niet aan de aannamen van het Scheffé-model, zoals genoemd in § 5.2.1, voldaan te zijn. Zowel de subtractiviteits- als de homogeniteitseis worden vaak geschonden en bovendien treden er herhaaldelijk orde-effecten op.

De oorzaak hiervan wordt reeds duidelijk bij bestudering van de diverse frequentiematrices. Deze vertonen vrijwel allen hetzelfde beeld : versie 3 (PSA-variabele puls) wordt zowel ten opzichte van versie 2 (PSA-vaste puls) als ten opzichte van versie 1 (PSA-standaard) maximaal slechter beoordeeld. Dit duidt op een identificeerbaarheid van versie 3, wat inderdaad bevestigd werd door uitspraken en commentaar achteraf van de proefpersonen. Dit is echter in tegenspraak met de aanname van het model, waarin de voorkeursoordelen als onafhankelijk verondersteld worden. We mogen de schaalwaarden die het Scheffé-programma als output levert daarom niet interpreteren. Blijkbaar is het verschil tussen versie 3 enerzijds en de versies 1 en 2 anderzijds te groot in vergelijking met het verschil tussen de versies 1 en 2 onderling.

Uit de uitspraken van proefpersonen bleek, dat t.g.v. de algehele spraakkwaliteit van versie 3, de proefpersonen er niet aan toe kwamen deze versie op natuurlijkheid te beoordelen. Daarom is een tweede experiment opgezet, waarbij de stimuli van een meer gelijke kwaliteit zijn.

### 5.2.3 Experiment 2

De versie PSA-variabele puls wordt nu niet meer bij het experiment betrokken. In plaats daarvan wordt de door het LVS-systeem gegenereerde spraakversie toegevoegd. Zoals reeds in het begin van dit hoofdstuk is vermeld, lijkt de LVS-spraak namelijk in eerste indruk van een vergelijkbare perceptieve kwaliteit als de PSA-standaard versie. Bovendien creëren we zo de mogelijkheid ons pitch-synchrone systeem perceptief te vergelijken met het niet-pitch-synchrone LVS-systeem. De LVS-spraakversies worden gegenereerd middels het bestaande LVS-systeem. Hierbij zijn de

programma's AAP (a-parameters), PCT en SYN gebruikt (Vogten,1985).

Twee van de in experiment 1 gebruikte sprekers (1 mannenstem en 1 vrouwenstem) worden in dit experiment niet meer meegenomen, aangezien deze sprekers na beluistering relatief slecht door een LPC-analyse-resynthese heen blijken te komen. Zodoende zijn dus van in totaal 6 sprekers (3 mannenstemmen, 3 vrouwenstemmen,zie appendix A) telkens weer 3 spraakstimuli beschikbaar, te weten :

- versie 1 : LVS
- versie 2 : PSA-standaard
- versie 3 : PSA-vaste puls

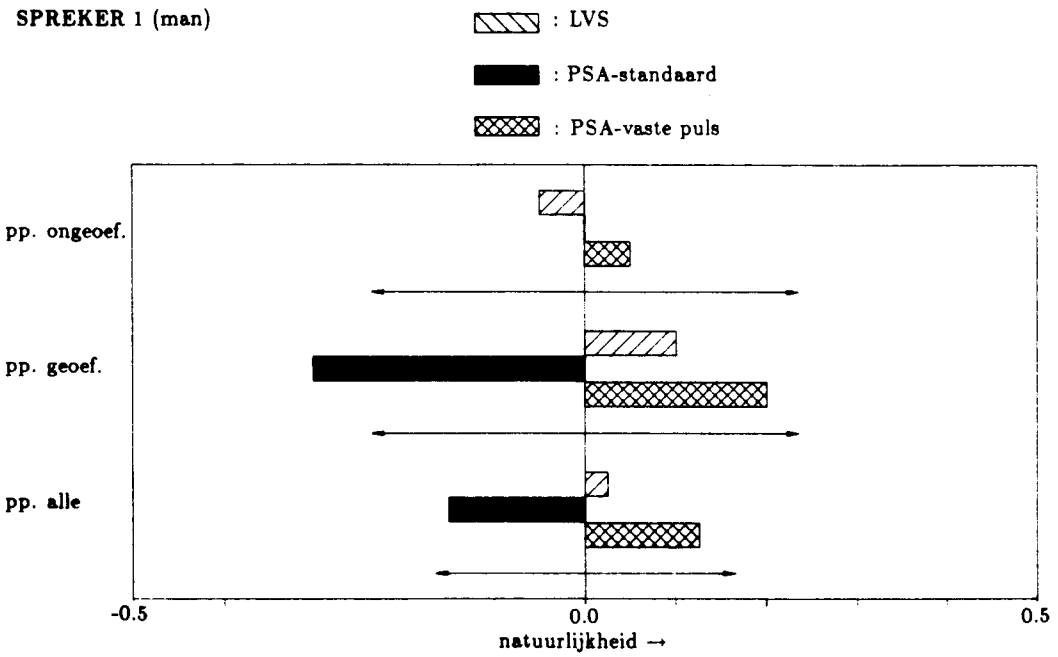
Voor het overige is dit experiment volkomen analoog aan experiment 1 (zelfde stimuluszin, zelfde proefpersonen, etc).

## Resultaten

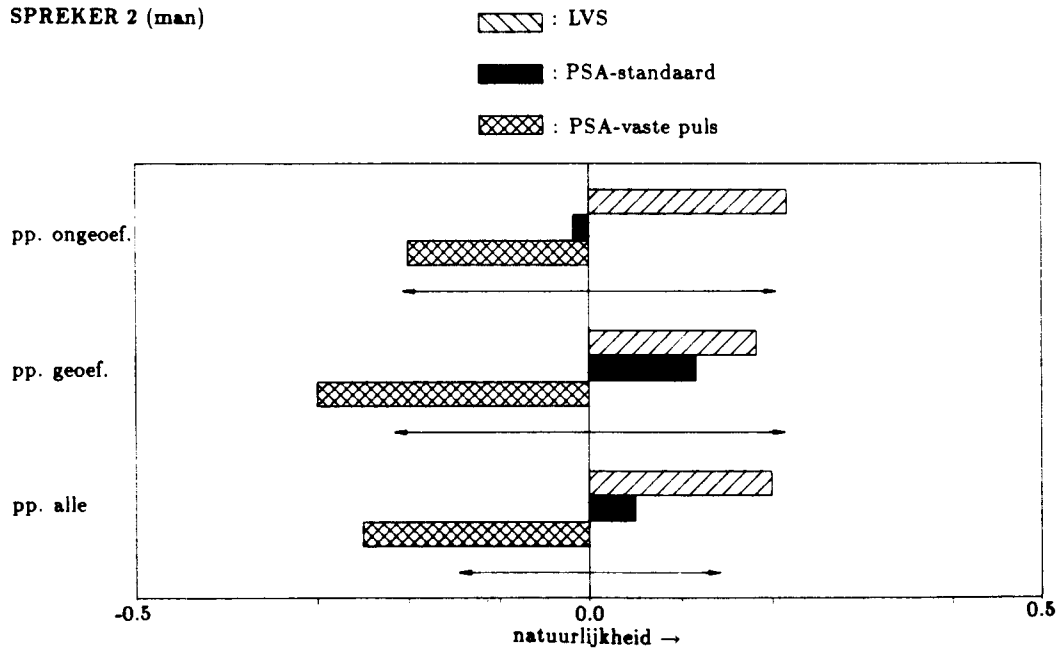
De resultaten van experiment 2 zijn grafisch weergegeven in de figuren 5.1 t/m 5.6. Iedere figuur bevat de resultaten van één spreker, uitgesplitst voor de ongeoefende proefpersonen, de geoefende proefpersonen en alle proefpersonen. De resultaten zijn weergegeven in de vorm van de berekende schaalwaarden behorende bij de drie verschillende versies. De dubbele pijlen geven telkens de grootte van de yardstick (het betrouwbaarheidsinterval) aan.

In alle gevallen bleek nu wel aan de aannamen van het Scheffé-model voldaan te zijn, zodat we de schaalwaarden onderling mogen vergelijken teneinde uitspraken te doen over mogelijke significante effecten. Kijken we allereerst naar de responsies van de ongeoefende proefpersonen, dan zien we dat bij 3 van de 6 sprekers geen significante verschillen tussen de 3 versies worden waargenomen. Bij de geoefende proefpersonen zijn, zoals verwacht, de verschillen in schaalwaarden duidelijk groter.

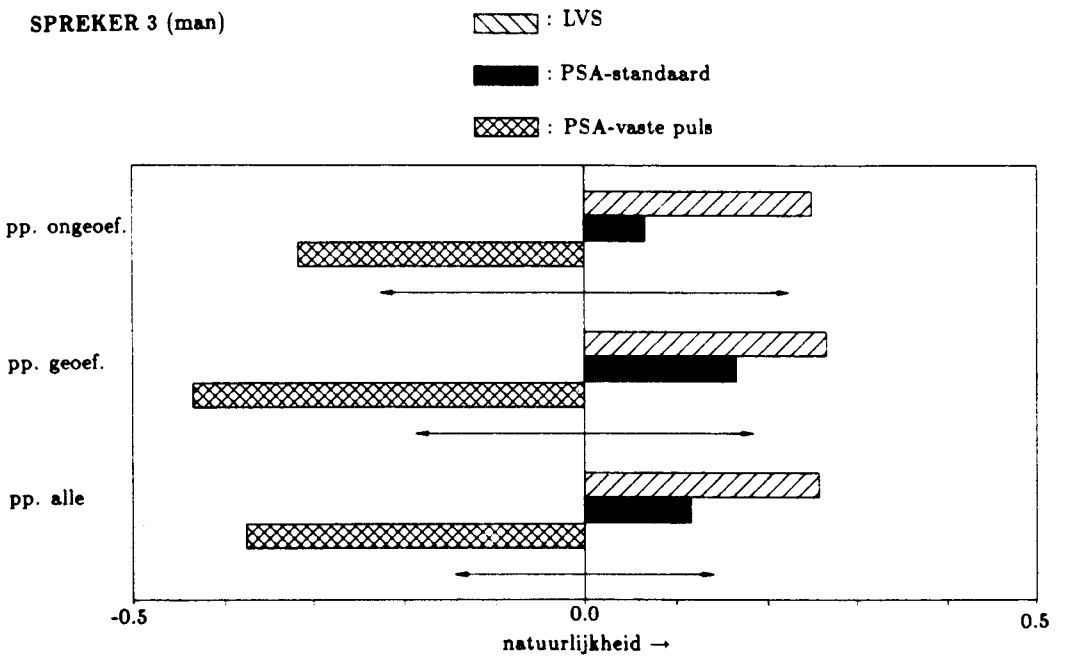
Kijken we naar significante verschillen tussen versie 2 (PSA-standaard) en versie 3 (PSA-vaste puls), dan blijkt bij de sprekers 2,3,4 en 6 versie 3 onnatuurlijker gevonden te worden dan versie 2. Dit verschil wordt meestal door de geoefende proefpersonen waargenomen. Voor spreker 1 daarentegen blijken de geoefende proefpersonen een voorkeur te geven voor versie 3 boven versie 2.



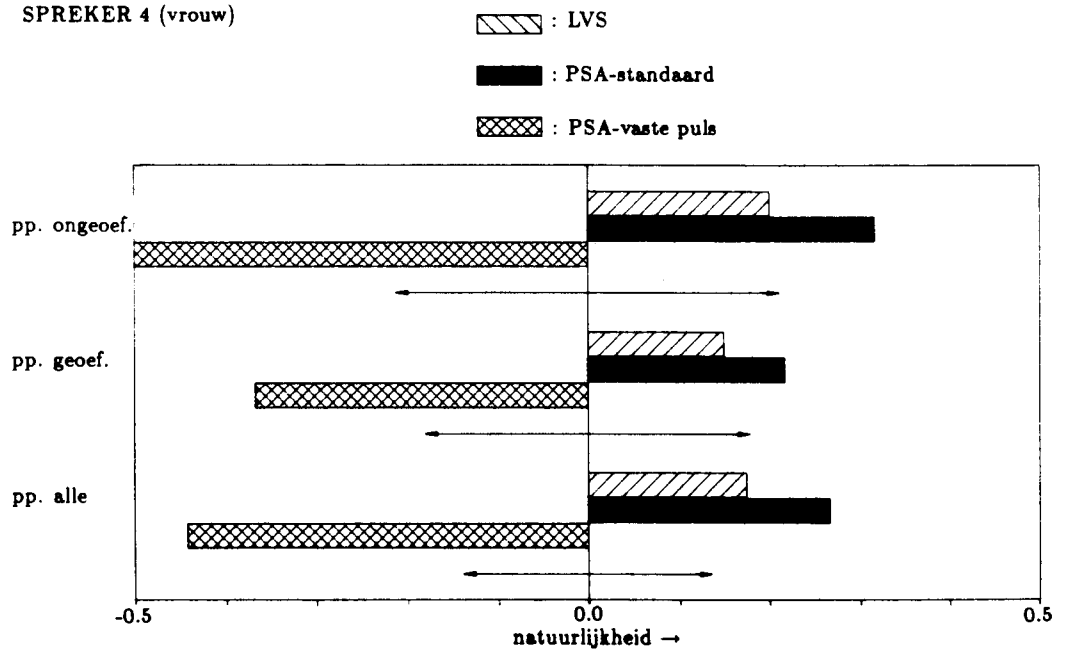

Figuur 5.1: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 1 (mannenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*



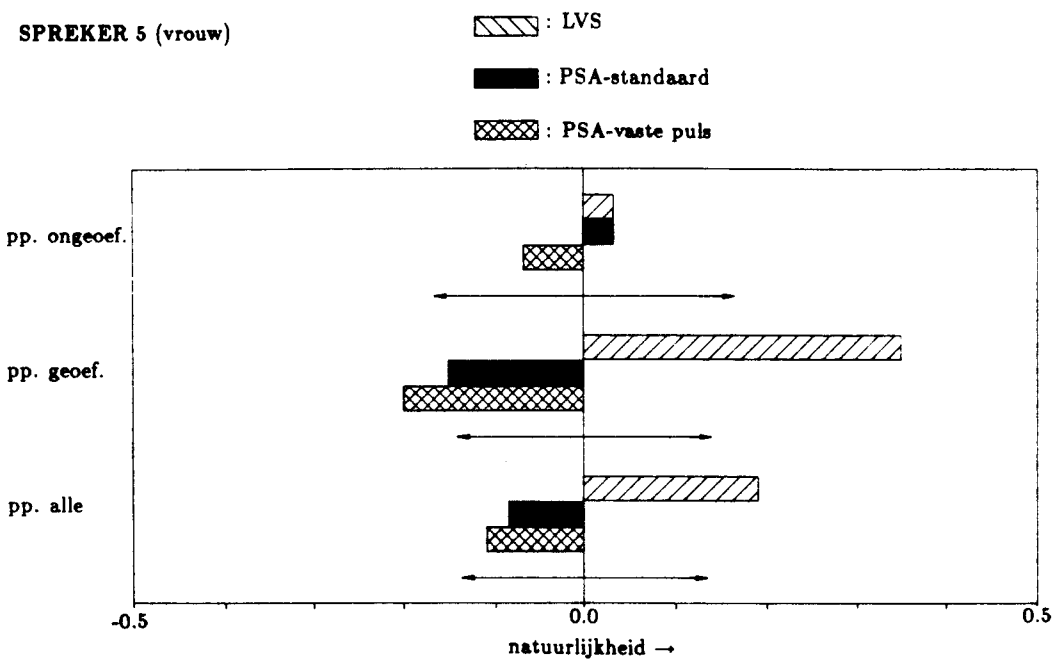

Figuur 5.2: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 2 (mannenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*



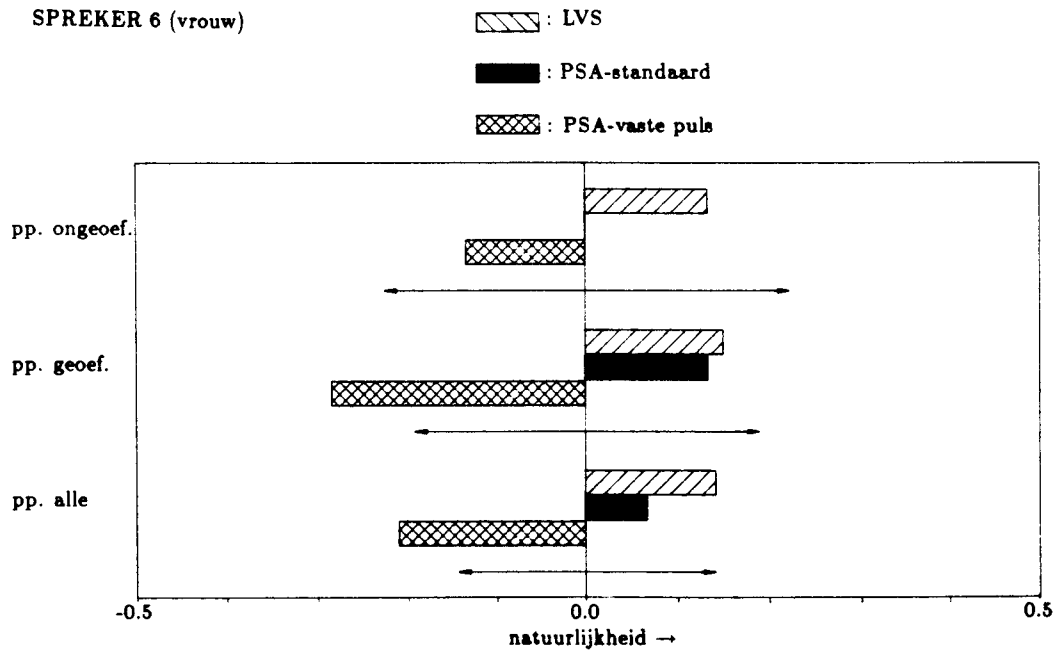

Figuur 5.3: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 3 (mannenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*




Figuur 5.4: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 4 (vrouwenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*




Figuur 5.5: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 5 (vrouwenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*




Figuur 5.6: *Schaalwaarden behorende bij de synthetische spraakversies LVS, PSA-standaard en PSA-vaste puls van spreker 6 (vrouwenstem), uitgesplitst voor alle pp., de geoefende pp. en de ongeoefende pp. (de pijlen geven het betrouwbaarheidsinterval aan).*

Verder zien we dat versie 3 bij alle sprekers, behalve bij spreker 1, onnatuurlijker beoordeeld wordt dan versie 1 (LVS).

Tenslotte kunnen we concluderen dat versie 1 en versie 2 niet significant verschillend worden waargenomen.

## 5.3 Conclusies en discussie

Uitgaande van de experimenten komen we nu tot de volgende perceptieve evaluatie van de in dit onderzoek ontwikkelde synthetische spraakversies.

Allereerst blijkt uit experiment 2 dat de standaard pitch-synchrone spraak perceptief van eenzelfde kwaliteit is als de niet-pitch-synchrone LVS spraak.

Experiment 2 geeft ons verder een beeld van het effect van het gebruik van de vaste excitatiefunctie (versie PSA-vaste puls) ten opzichte van het gebruik van een eenheidsimpuls als excitatiefunctie (versie PSA-standaard). Het blijkt dat het effect verschillend is voor de verschillende sprekers : bij één mannelijke spreker ontstaat een verbetering in de natuurlijkheid, terwijl bij vier andere sprekers (2 mannen,2 vrouwen) een verslechtering optreedt. We kunnen dus concluderen dat het effect van het gebruik van een minder geïdealiseerde excitatiefunctie dan een eenheidsimpuls, niet vanzelfsprekend gegeneraliseerd kan worden voor alle sprekers. Eerder in de literatuur vermelde resultaten (Rosenberg,1971 / Sambur et al.,1978) toonden een vergelijkbare verbetering van de natuurlijkheid op basis van een modellering van het bronsignaal aan. Deze resultaten hadden echter slechts betrekking op één (mannelijke) spreker. Nú blijkt dus dat dit resultaat niet voor alle sprekers geldt. Wellicht is het verschillende resultaat voor de diverse sprekers te verklaren door het feit dat er één keuze is gemaakt voor de parameterwaarden, die de excitatiefunctie vastleggen. Met name de open-dichttijd verhouding R/F is in navolging van Rosenberg(1971) gelijk gekozen aan 40% / 16%. Wellicht bestaat er echter voor iedere spreker afzonderlijk een optimale keuze voor deze parameters.

Toch moeten we ons afvragen of het gebruik van de vaste excitatiepuls inderdaad slechts bij één van de zes sprekers tot een verbetering van de natuurlijkheid leidt. Het informeel beluisteren van de versie PSA-vaste puls door de bij dit onderzoeksproject betrokken personen, gaf namelijk voor meerdere sprekers de indruk, dat deze versie een vollere, donkerdere

klank heeft dan de versie PSA-standaard, wat als een verbetering van de natuurlijkheid ervaren werd. Het feit dat dit niet in dié mate in het experiment tot uiting komt, zou veroorzaakt kunnen worden door de reeds eerder genoemde niet-eenduidigheid van het criterium 'natuurlijkheid', waardoor de proefpersonen andere criteria in hun oordeel kunnen betrekken, dan verwacht wordt. Zo kunnen proefpersonen een stimulus als minder natuurlijk beoordelen t.g.v. slechts één analysefout ('tik', 'plop' e.d.), terwijl bij de interpretatie van de responsies er van uitgegaan wordt, dat de stimuli op een 'algehele natuurlijkheid' beoordeeld zijn.

Een tweede oorzaak zou gelegen kunnen zijn in de relatief toch nog vrij slechte kwaliteit van LPC-gesynthetiseerde spraak, waardoor proefpersonen wellicht niet aan een beoordeling van natuurlijkheid toekomen.

Een ander effect wat nog een rol zou kunnen spelen, is het feit dat met name de geoefende proefpersonen op basis van gewenning een intuïtieve voorkeur voor LVS-spraak kunnen hebben. Wat dat betreft geven ongeoefende luisteraars wellicht een spontaner oordeel. Het probleem bij het gebruik van ongeoefende proefpersonen is echter weer dat deze minder gauw verschillen horen. De keuze van de soort proefpersonen blijft dus in dit soort experimenten een moeilijke kwestie.

Uit experiment 1 blijkt tenslotte dat de algehele kwaliteit van de versie PSA-variabele puls in die mate te wensen over laat, dat het gebruik van de variabele excitatiefunctie zich niet uit in een betere natuurlijkheid. Tijdens de ontwikkeling van een variabele parametrisering ontstond echter bij het beluisteren van bepaalde *korte* uitgesneden spraakfragmenten toch de indruk dat er sprake was van een duidelijk hoorbare verbetering. Bij het beluisteren van de gehele spraakuiting blijken optredende analyse en/of parametriseringsfouten deze verbeteringen echter teniet te doen. In dit verband zou het wellicht bij een perceptieve evaluatie beter zijn allereerst gebruik te maken van kortere stimuli dan een gehele zin.

# Hoofdstuk 6

# Conclusies

Resumerend kunnen we op basis van het in dit rapport bechreven onderzoek tot de volgende conclusies komen.

Allereerst is een pitch-synchroon analyse-resynthese systeem ontwikkeld, waarvan de perceptieve kwaliteit overeenkomt met het reeds bestaande IPO-LVS systeem. Met behulp van dit systeem is de mogelijkheid gecreëerd om het effect van het gebruik van verschillende excitatiefuncties te onderzoeken. Daarnaast zou dit systeem tevens gebruikt kunnen worden voor andere doeleinden binnen het spraakonderzoek, waarbij het noodzakelijk is bewerkingen en/of manipulaties per pitch-periode uit te voeren.

Het gebruik van een minder geïdealiseerde excitatiefunctie dan een eenheidsimpuls in de vorm van een *vaste* parametrisering van het door de stembanden opgewekte bronsignaal, blijkt de natuurlijkheid van synthetische spraak te kunnen verbeteren. Uit dit onderzoek blijkt verder dat het effect van een dergelijke vaste parametrisering niet voor alle sprekers generaliseerbaar is. Dit vormt een belangrijk resultaat in het kader van eerdere in de literatuur vermelde resultaten (Rosenberg,1971 / Sambur et al.,1978), waarbij een verbetering in natuurlijkheid gerapporteerd werd op basis van één (mannelijke) spreker. Uit het in dit rapport bechreven onderzoek blijkt dit resultaat dus niet vanzelfsprekend geldig te zijn voor alle sprekers. Een mogelijke verklaring van het gevonden verschil in effect van de vaste parametrisering voor diverse sprekers zou kunnen liggen in de keuze, die gemaakt is ten aanzien van de parameterwaarden van het gebruikte parametriseringsmodel. Wellicht bestaat er voor deze parameterwaarden een optimale keuze, die voor iedere spreker verschillend is.

Voor wat betreft de implementatie van de excitatiefunctie beschreven door variabele parameters, moeten we het volgende concluderen. De gebruikte parametrisering in de vorm van een stilering van het signaal dat ontstaat na inverse-filtering van het natuurlijke spraaksignaal, blijkt tot een algemene spraakkwaliteit te leiden, die in die mate te wensen over laat, dat het gebruik van de variabele excitatiefunctie niet in een betere natuurlijkheid tot uiting komt. Toch ontstond bij de ontwikkeling van de variabele parametrisering van de excitatiefunctie de indruk, dat bepaalde korte fragmenten spraak duidelijk natuurlijker klonken.

Het lijkt daarom zinvol nader onderzoek te verrichten naar een juiste parametrisering van de excitatiefunctie op basis van het d.m.v. inverse-filtering verkregen signaal. Met name zal verder onderzoek gericht moeten zijn op een beter begrip van de betekenis van dit signaal, waarbij de relatie met de glottale puls centraal zal moeten staan, zodat tevens een meer directe aansluiting met de literatuur mogelijk wordt.

# Referenties


[1] Atal B.S. & Hanauer S.L. (1971),
"Speech analysis and synthesis by linear prediction of the speech wave",
*Journal of the Acoustical Society of America 50, 637-655.*

[2] Brady P.T. (1968),
"Equivalent Peak Level : a Threshold-Independent Speech-Level-Measure",
*Journal of the Acoustical Society of America 44,695-699.*

[3] Damen G.H.T. & Ellermann H.H. (1988),
"An analysis of variance for experiments with paired comparisons : introduction and application",
*IPO rapport no. 663, Institute for Perception Research Eindhoven.*

[4] Damen G.H.T. (1988),
"Micro-intonatie in kunstmatige spraak",
*IPO rapport no. 664, Institute for Perception Research Eindhoven.*

[5] Fant G. (1960),
"Acoustic theory of speech production",
*Mouton, 's Gravenhage.*

[6] Flanagan J.L. (1972),
"Speech analysis, synthesis and perception",
*Springer-Verlag, Berlijn (2nd edition).*

[7] Fujisaki H. & Ljungqvist M. (1986),
"Proposal and evaluation of models for the glottal source waveform",
*ICASSP 86, 1605-1608.*

[8] Fujisaki H. & Ljungqvist M. (1987),
"Estimation of voice source and vocal tract parameters based on ARMA-analysis and a model for the glottal source waveform",
*IEEE, 637-640.*

[9] Hart J.'t, Nooteboom S.G., Vogten L.L.M. en Willems L.F. (1981-82),
"Manipulaties met spraakgeluid",
*Philips Techn. Rev. 40, 108-119.*

[10] Hemert J.P. van (1987),
"The voiced and unvoiced amplitude in speech synthesis",
*IPO rapport no. 595, Institute for Perception Research Eindhoven.*

[11] Hermes D.J. (1986),
"Measurement of pitch by subharmonic summation",
*IPO Annual Progress Report 21, 24-33, Institute for Perception Research Eindhoven.*

[12] Holmes J.N. (1973),
"The influence of glottal waveform on the naturalness of speech from a parallel formant-synthesizer",
*IEEE Trans. Audio and Electro-Acoust. Au-21, 298-305.*

[13] Holmes J.N. (1976),
"Formant excitation before and after glottal closure",
*ICASSP 76, 39-42.*

[14] Hovelynck I.C.M. (1985),
"Een overzicht van subjectieve spraakkwaliteitstests",
*Notitie 85 nw/232, PTT-dr.Neher Laboratorium.*

[15] Makhoul J. (1975),
"Linear prediction : a tutorial review",
*Proceedings IEEE 63, 561-580.*

[16] Markel J.D. & Gray A.H. (1976),
"Linear prediction of speech",
*Springer-Verlag, Berlijn.*

[17] Nooteboom S.G. & Cohen A. (1976),
"Spreken en verstaan",
*Van Gorcum, Assen.*

[18] Rosenberg A.E. (1971),
"Effect of glottal pulse shape on the quality of natural vowels",
*Journal of the Acoustical Society of America 49, 583-590.*

[19] Rothenberg M. (1973),
"A new inverse filtering technique for deriving the glottal air flow waveform during voicing",
*Journal of the Acoustical Society of America 53, 1632-1645.*

[20] Sambur M.R., Rosenberg A.E., Rabiner L.R., McGonegal C.A. (1978),
"On reducing the buzz in LPC-analysis",
*Journal of the Acoustical Society of America 63, 918-924.*

[21] Scheffé H. (1952),
"An analysis of variance for paired comparisons",
*Journal of the Statistical Society of America 47, 381-400.*

[22] Vogten L.L.M. (1983),
"Analyse,zuinige codering en resynthese van spraakgeluid",
*proefschrift Technische Universiteit Eindhoven.*

[23] Vogten L.L.M. (1985),
"LVS-speech processing programs on IPO-VAX 11/780" *IPO handleiding no. 67, Institute for Perception Research Eindhoven.*

[24] Vogten L.L.M et al. (1988),
Syllabus behorend bij het college Spraaktechnologie 0H050,
*IPO rapport no. 649, Instituut voor Perceptie Onderzoek Eindhoven.*

[25] Witten J.H. (1982),
"Principles of computer speech",
*Academic Press, London.*

# Nawoord

Hierbij wil ik allereerst een woord van dank richten aan prof.dr. H. Bouma, voor de mogelijkheid tot het uitvoeren van dit afstudeerproject op het Instituut voor Perceptie Onderzoek, alsmede voor zijn getoonde belangstelling.

Daarnaast een woord van dank aan Lei Willems en Berry Eggen, die mij gedurende dit onderzoek begeleid hebben. In het bijzonder wil ik Berry bedanken voor zijn spontane betrokkenheid en ondersteuning in alle fasen van het onderzoek en voor de vele vruchtbare discussies die ik met hem mocht hebben.

Verder wil ik Gerd Damen bedanken voor zijn behulpzaamheid bij het experimentele gedeelte van dit afstudeerproject.

Ook bedank ik alle andere medewerkers cq. studenten van het IPO, die ieder op hun eigen manier aan dit onderzoek hebben bijgedragen.

En last, but sure not least, Christel, zonder wiens jarenlange steun en inspiratie deze eindfase nooit bereikt zou zijn !

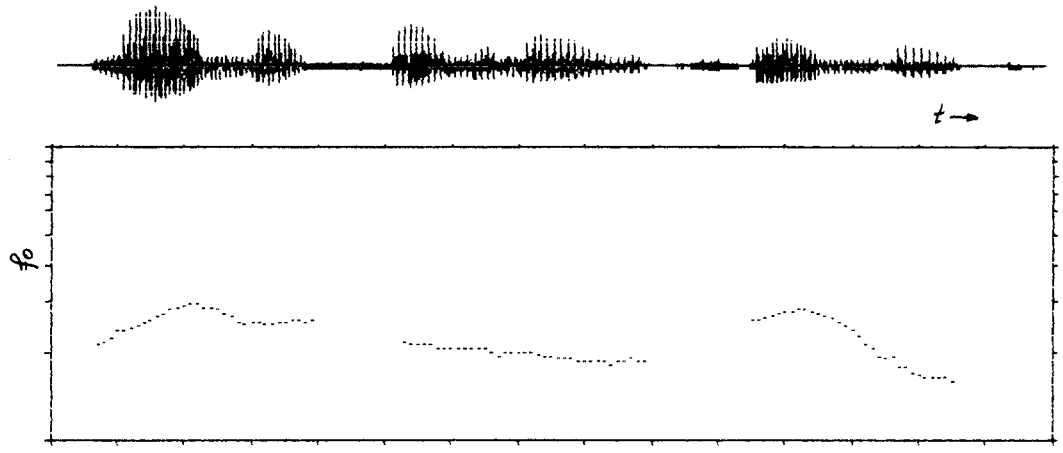
# Appendix A

## Lijst van in de experimenten gebruikte sprekers

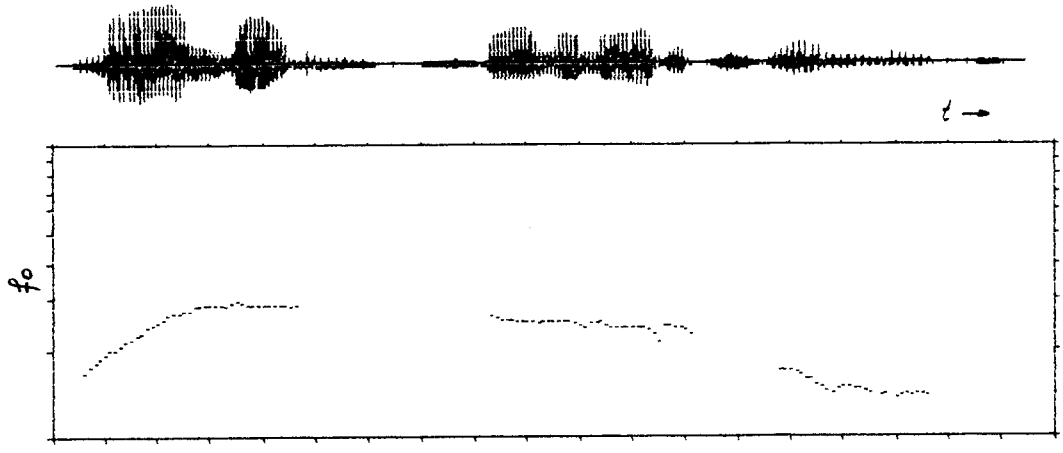
| spreker | geslacht | experiment |
|---|---|---|
| 1. Ben Elsendoorn | M | 1,2 |
| 2. Jan v. Hemert | M | 1,2 |
| 3. Hans 't Hart | M | 1,2 |
| 4. Roel Smits | M | 1 |
| 5. Christel Pan | V | 1,2 |
| 6. Anja v.d. Water | V | 1,2 |
| 7. Joyce Hofhuis | V | 1,2 |
| 8. Ingrid Borgharts | V | 1 |

In de 3e kolom is telkens weergegeven in welk experiment de betreffende spreker is gebruikt.

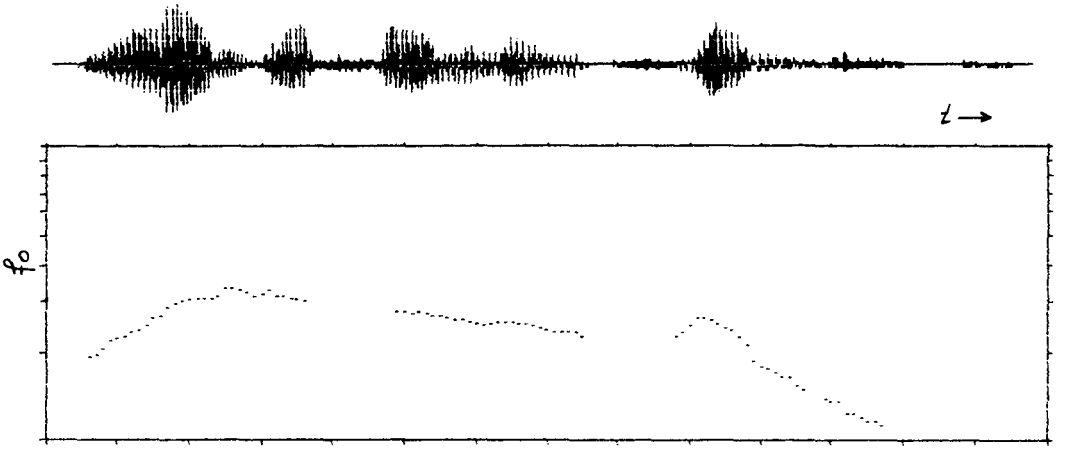
In de figuren A.1 t/m A.8 zijn voor iedere spreker zowel de golfvorm als het verloop van de grondtoon weergegeven voor de in de experimenten gebruikte spraakuiting "Maandag gaan we naar het zwembad".



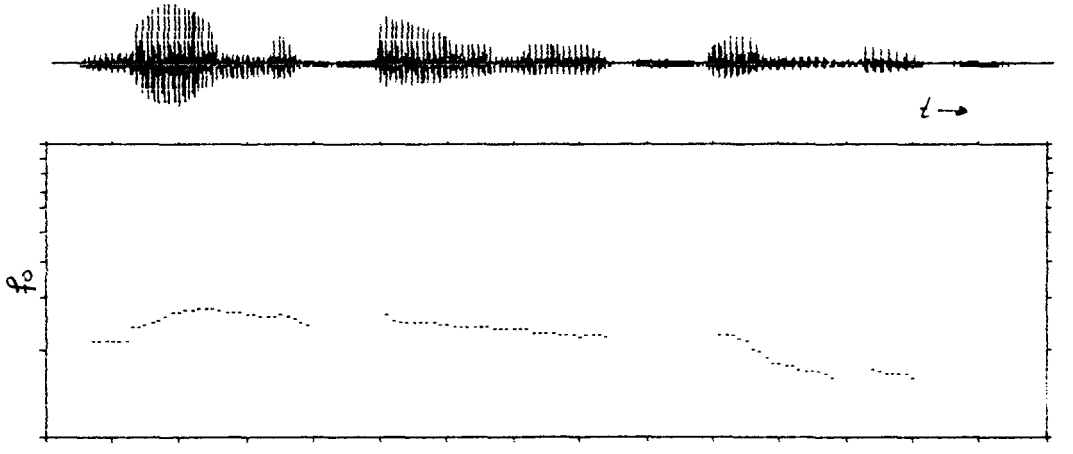

Figuur A.1: *Spreker no.1*



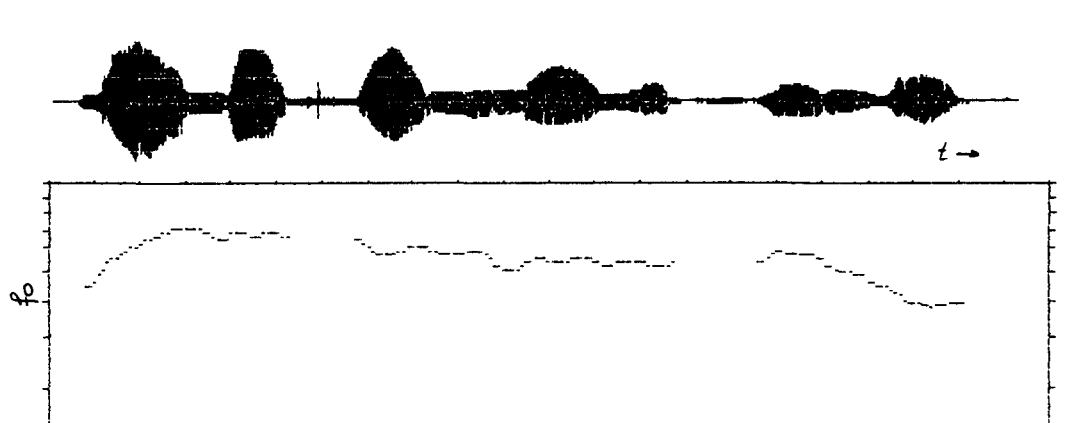

Figuur A.2: *Spreker no.2*



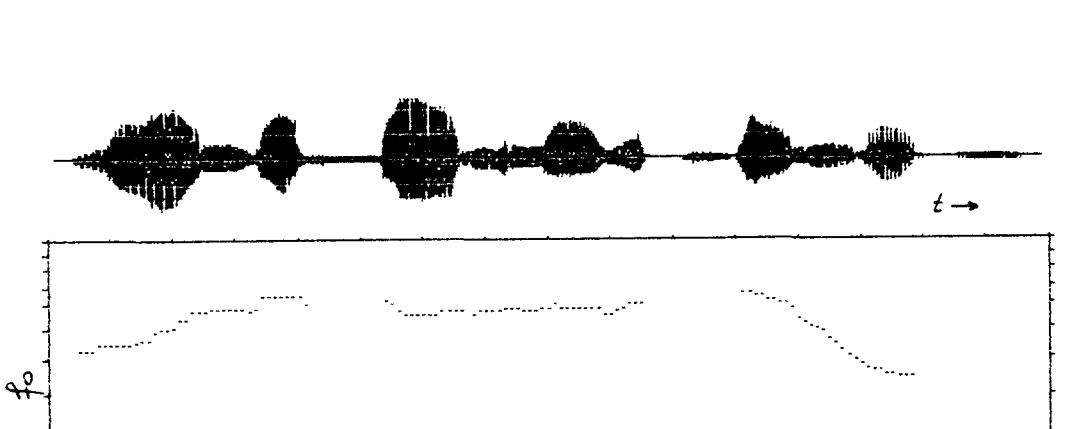

Figuur A.3: *Spreker no.3*



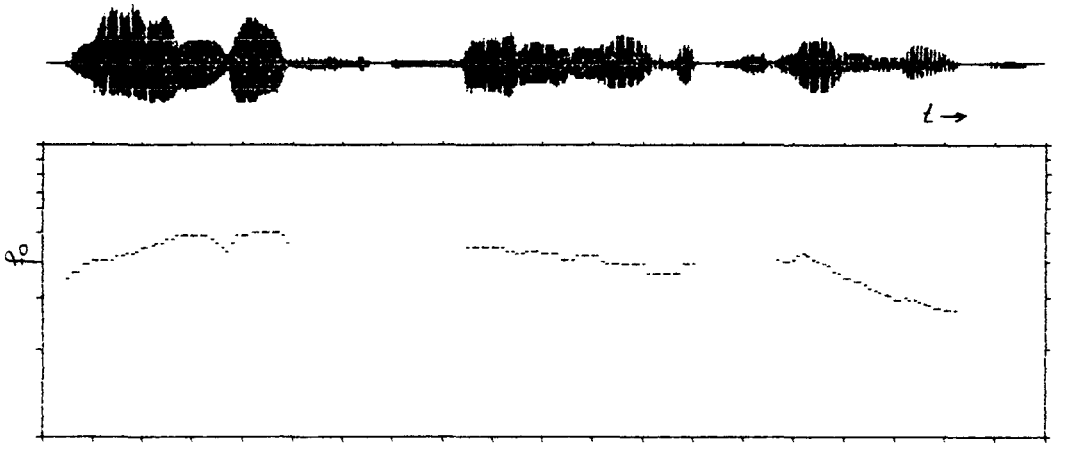

Figuur A.4: *Spreker no.4*



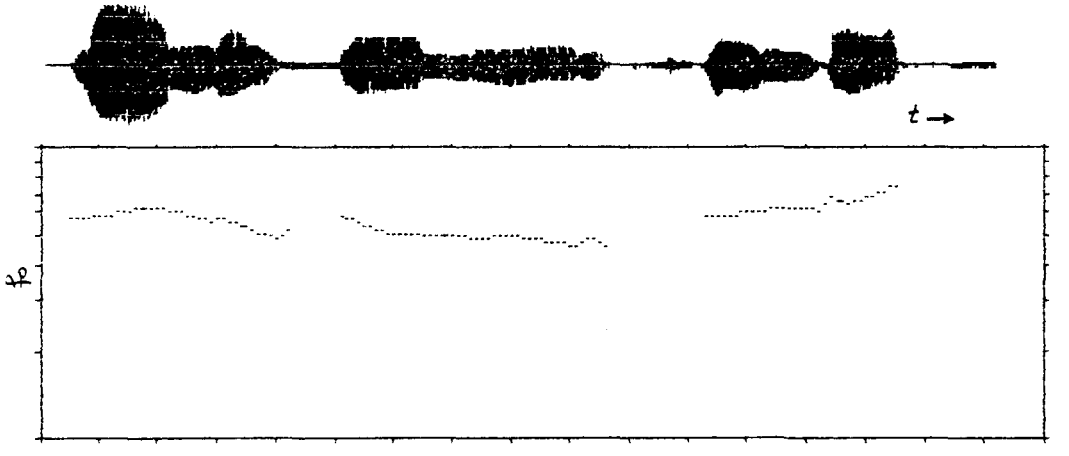

Figuur A.5: *Spreker no.5*




Figuur A.6: *Spreker no.6*




Figuur A.7: *Spreker no.7*




Figuur A.8: *Spreker no.8*

# Appendix B

## Voorbeeld van het bij de experimenten gebruikte antwoordformulier :

**Naam proefpersoon :**

**Geoefend spraakluisteraar ? (J/N) :**

**Datum experiment :**

**Antwoordformulier :**

lange piep
3 proefstimuli spreker 1
lange piep
lange piep

| | | 1 ← | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep
lange piep
3 proefstimuli spreker 2
lange piep
lange piep

| | | 1 ← | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep

*- zie volgende pagina -*

lange piep
3 proefstimuli spreker 3
lange piep
lange piep

| | 1 ← | | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep
lange piep
3 proefstimuli spreker 4
lange piep
lange piep

| | 1 ← | | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep
lange piep
3 proefstimuli spreker 5
lange piep
lange piep

| | 1 ← | | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep

*- zie volgende pagina -*

lange piep
3 proefstimuli spreker 6
lange piep
lange piep

| | | 1 ← | | → 2 | |
|---|---|---|---|---|---|
| paarvergelijking 1 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 2 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 3 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 4 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 5 | -2 | -1 | 0 | 1 | 2 |
| paarvergelijking 6 | -2 | -1 | 0 | 1 | 2 |

lange piep

**einde experiment**

Vraag : kon u de stimuli identificiëren ?
(d.w.z. wist u bij het aanhoren van een bepaalde paarvergelijking, met welke 2 versies u te maken had ?)

Eventuele verdere opmerkingen :

# Appendix C

## Voorbeeld van het bij de experimenten gebruikte instructieformulier :

**Instructie :**

In het volgende experiment krijgt u een aantal synthetische spraakstimuli te horen. Achtereenvolgens worden er van 6 sprekers telkens 3 verschillende versies aangeboden. Doel van het experiment is het per spreker rangschikken van deze 3 versies naar mate van natuurlijkheid.
Hiertoe krijgt u per spreker de 3 verschillende versies paarsgewijs aangeboden, waarbij u op het antwoordformulier kunt aangeven welke versie voor u het meest natuurlijk overkomt. U kunt uw oordeel hierbij in de volgende gradaties aangeven :

-2 : versie 1 duidelijk natuurlijker.
-1 : versie 1 iets natuurlijker.
0 : beide versies even natuurlijk.
1 : versie 2 iets natuurlijker.
2 : versie 2 duidelijk natuurlijker.

De gebruikte stimuluszin luidt : "Maandag gaan we naar het zwembad".

De stimuli worden als volgt aangeboden :

- 3 proefstimuli spreker 1 : elke versie 1 maal in willekeurige volgorde.
- 6 paarvergelijkingen spreker 1.

Idem voor de andere 5 sprekers.

Bij de proefstimuli hoeft u geen antwoord te geven, ze zijn bedoeld ter oriëntatie.
Per paarvergelijking hebt u 4 seconden bedenktijd.
Tussen de paarvergelijkingen zitten steeds korte piepjes.

# Appendix D

## Lijst van ontwikkelde pitch-synchrone software-programmatuur

- **PAN** : voert een Pitch-ANalyse uit op de stemhebbende gedeelten van een N-spraakfile, waarbij de periodes in het spraaksignaal bepaald worden uit het verloop van de locale covariantiefout in het signaal. Stemloze gedeelten worden verdeeld in "periodes" van 100 samples.
- **EDTCOL** : Programma waarmee de pitchpoints, gevonden met het programma PAN interactief gecorrigeerd kunnen worden.
- **PSAN** : voert een Pitch-Synchrone-ANalyse uit op een N-spraakfile.
- **PSS** : voert een Pitch-Synchrone-Synthese uit met als excitatiefunctie een deltapuls of een puls volgens Fujisaki-Ljungqvist model.
- **PSSSTY** : Pitch-Synchrone-Synthese met het gestileerde 2 maal geïntegreerde residusignaal als excitatiefunctie.
- **PSSRES** : Pitch-Synchrone-Synthese met residusignaal als excitatie (levert weer de originele spraak op).
- **RESIDU** : pitch-synchrone berekening van residusignaal. Berekent tevens het 1 maal en 2 maal geïntegreerde residusignaal.
- **STYRES** : stilering van 2 maal geïntegreerde residusignaal.