L.P.C.-analyse en formantsynthese van spraak

P 308

L.P.C.-ANALYSE EN FORMANTSYNTHESE VAN SPRAAK

Ir. L. F. Willems
Instituut voor Perceptie Onderzoek, Eindhoven

In het kort worden hier de principes van een betrekkelijk nieuwe analyse-synthese-methode van spraak beschreven, met welke methode een goede kwaliteit van de gereproduceerde spraak is te bereiken (Linear Predictive Coding). Ten behoeve van het fonetisch onderzoek geschiedt het synthetiseren met behulp van formanten.

1) INLEIDING

Het vinden van een representatie van spraaksignalen in een aantal slechts langzaam variërende parameters is van groot belang voor een aantal toepassingen in het spraakonderzoek. De mens produceert een gering aantal (5 à 10) spraakklanken per seconde, want de articulatoren en dus ook het spraakkanaal kunnen slechts met een beperkte snelheid bewegen. Ook het menselijk gehoororgaan, langs welke weg de spraak bij de mens binnenkomt , is beperkt wat betreft het verwerken van het aantal verschillende klanken per seconde. Uit dit soort van overwegingen is het plausibel te maken dat de informatie-inhoud van bijvoorbeeld PCM-gecodeerde spraak van 48000 bits/sec (6000 Hz aftastfrequentie en 8 bits per sample) aanzienlijk te reduceren is. Als ondergrond komt men tot ongeveer 60 bits/sec. (stelt men dat er 64 verschillende spraakklanken zijn, die in 6 bits zijn te coderen, en stelt men dat 10 spraakklanken per seconde worden geproduceerd, dan geeft dit 60 bits/sec.).

Toepassingen van deze in de praktijk echter niet zo dramatische informatie-reductie bij spraaksignalen, liggen op het terrein van de spraakherkenning, van spraakopslag t.b.v. zogenaamde voice response en van de overdracht van spraak. Op het gebied van de spraaktransmissie noemt men deze toepassingen: vocoders. Hoewel het idee van de vocoder uit de dertiger jaren stamt en in de tijd daarna er veel uitvoeringsvormen van vocoders zijn voorgesteld (SCHRÖDER 1966), zijn ze nooit op grote schaal toegepast, ofwel de bereikte bandbreedtereductie was te gering ofwel de kwaliteit van het uiteindelijke spraaksignaal was te slecht. De laatste jaren is de belangstelling voor vocoders weer toegenomen, voornamelijk door de grote vlucht van de digitale technieken. Ook biedt de digitale vorm goede perspectieven voor geheime coderingen tijdens de spraakoverdracht.

Een van de meer complexe, maar toch in de praktijk uitvoerbare analyse-synthese-technieken is de zog. Linear Predictive Coding: LPC afgekort. (ITAKURA 1969, ATAL and HANAUER 1971). In het nabije verleden is aangetoond dat enerzijds zeer goede kwaliteit van de spraak is te bereiken (ATAL) en anderzijds een grote bandbreedtereductie tot beneden 1000 bits/sec is te behalen (SAMBUR 1975, KANG and COULTER 1970).

De motivatie voor ons om aan dergelijke analyse-synthese-systemen te werken zijn de toepassingen ervan bij het fonetisch onderzoek. Bij studies omtrent de waarneming van spraakklanken heeft de experimentator vaak behoefte aan stimuli, die volgens bepaalde voorschriften zijn gemaakt of gevarieerd. Het maken van dergelijke stimuli kan geschieden m.b.v. analyse-synthese-systemen van spraak. Wil men de waarneming van de intonatie van zinnen bestuderen, dan moet men de luisteraar (proefpersoon) zinnen kunnen voorspelen, waarin de intonatie systematisch wordt gevarieerd. Daartoe wordt de toonhoogte (dat is de grondfrequentie van het brongeluid) bij het syntheseproces vervangen door een kunstmatig opgewekt toonhoogteverloop, zoals dat door de experimentator wordt gevraagd.

In deze bijdrage wordt een korte beschrijving gegevens van deze LPC analyse-synthesetechniek. In de fonetiek is echter een beschrijving van spraakklanken in termen van de zog. formanten gebruikelijk. Formanten zijn de resonantiefrequenties van het mondkanaal. Bij een neutrale klinker van een mannenstem liggen de formanten resp. bij: 500 Hz eerste formant, 1500 Hz tweede formant, 2500 Hz derde formant. De LPC techniek levert een goede mogelijkheid de analysegegevens om te rekenen naar een formantenbeschrijving. Zodoende kan het syntheseproces gebruik maken van deze in de fonetiek gekende formanten.

2) LINEAR PREDICTIVE CODING

Hierbij gaat men uit van een productiemodel van spraak dat bestaat uit een bron, die ofwel een periodieke puls ofwel ruis produceert en een lineair filter dat alleen polen bevat (fig.1a).

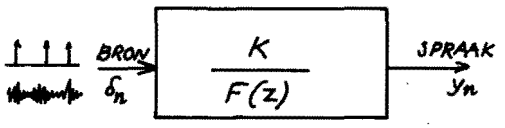

Fig. 1a. Spraakproductiemodel.

Hoewel de overdrachtsfunctie van het mondkanaal in sommige gevallen ook nulpunten bevat en hoewel het brongeluid, dat door de stembanden wordt geproduceerd, niet pulsvormig van aard is, is het de veronderstelling dat met het genoemde model het spraakproductieproces voldoende nauwkeurig is te benaderen. Het filter dat alleen polen bevat is in fig. 1b voorgesteld als een recursief filter:

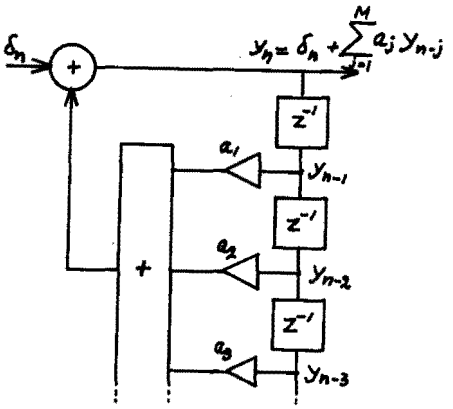

Fig. 1b. Spraakproductiemodel als recursief filter.

ook output spraaksample S_n kan worden geschreven als de som van de input δ_n en een lineaire combinatie van een aantal spraaksamples S_{n-j} uit het verleden:

$$S_n = \delta_n + \sum_{j=1}^{m} a_j S_{n-j} \qquad (1)$$

De lineaire combinatie van een aantal spraak samples uit het verleden is op te vatten als een voorspelling; vandaar de naam Linear Predictive Coding. De parameter M bepaalt de orde van het filter en is het aantal polen dat bij de benadering wordt gebruikt. M ligt in de praktijk tussen 10 en 15. De coëfficienten a_j voor $5 = 1, 2, \ldots\ldots, M$ bepalen de overdrachtsfuncties en ze vormen samen met nog enkele andere parameters (Fo en stemhebbend- stemloos parameter) een beschrijving van het spraaksignaal op een bepaald moment. De waarden van deze coëfficienten a_j veranderen betrekkelijk langzaam en worden in de praktijk 100 of 50 per seconde bepaald.

Bij de bepaling van de coëfficienten a_j gaat men uit van een stuk spraak van bijvoorbeeld 25 ms, overeenkomend met 250 samples als met 10 kHz wordt afgetast. Dit aantal noemen we N. De veronderstelling is dat het spraakkanaal gedurende dat tijdsinterval als stationair is te beschouwen. Het stelsel coëfficienten $\{a_j\}$ wordt uit dit stuk spraak door een minimaliseringsprocedure bepaald. Dit geschiedt als volgt: Door de predictor (het filter) kan een spraaksample worden voorspeld. Daarbij treedt een fout op t.o.v. het spraaksample S_n in het analyse-interval:

$$E_n = S_n - \sum_{j=1}^{M} a_j S_{n-j} \qquad (2)$$

Door het minimaliseren van de gemiddelde kwadratische fout: $\{E_n^2\}$ gem worden de coëfficienten $\{a_j\}$ gevonden.

$$\frac{\partial}{\partial a_k}\{(S_n - \sum_{j=1}^{M} a_j S_{n-j})^2\}_{gem} = 0 \qquad (3)$$

voor $K = 1,2,\ldots\ldots,M$.

Hieruit volgt een stelsel vergelijkingen:

$$\sum_{j=1} a_j R_{|K-j|} = R_K \text{ voor } K=1,2,\ldots\ldots,M \qquad (4)$$

waarin

$$R_K = \sum_{n=0}^{N-1-k} S_n S_{n+k} \qquad (5)$$

Deze laatste grootheden R_k zijn de auto-correlatie-coëfficienten van het stuk spraak in het analyse interval. De matrix die in dit stelsel vergelijkingen voorkomt is van een speciale vorm, waardoor het stelsel vergelijkingen recursief en snel is op te lossen.

Aan de parameters nodig om het spraaksignaal compleet te beschrijven ontbreken nog enkele (fig. 1a), nl. de amplitude van het signaal, het gegeven of ruis of periodiek signaal als bron moet dienst doen en in het geval het bronsignaal periodiek is, is het nodig de herhalingsfrequentie ervan te kennen. Het meten van deze herhalingsfrequentie is een probleem waarop we later nog terugkomen.

De voorspellen (het filter) die bij de bepaling van de coëfficient en aj wordt gehanteerd wordt ook wel invers filter genoemd (fig. 2). De output van het filter als men de spraak op de input zet is immers (in de zin der kleinste kwadraten) geminimaliseerd. Het inverse filter is

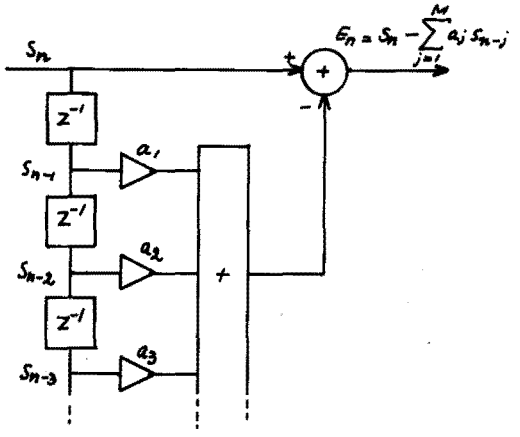

Fig. 2. Voorspel-fout bij de analyse.

dan ook op te vallen als een bewerking op het ingangssignaal welke naar vermogen (aantal polen M) het spectrum probeert glad te strijken.

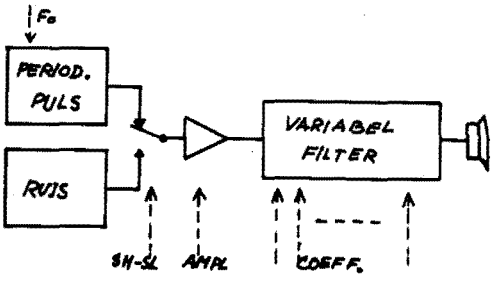

Fig. 3. Synthese-schema.

Het schema dat wordt gebruikt om de spraak m.b.v. deze beschrijving weer te resynthetiseren (fig. 3) correspondeert vanzelfsprekend met het spraakproduktiemodel, dat ten grondslag ligt aan deze methode: De stemhebbend- stemloosschakelaar laat ofwel ruis ofwel een periodieke puls met herhalingsfrequentie Fo toe tot de amplitude modulator en daarna aan het variable filter.

De structuur van het variabelefilter kan allerlei vormen aannemen. De eenvoudigste structuur is een recursief filter met de coëfficienten a_j zoals in figuur 1b. Dat is een goede manier voor de synthese van spraak en levert goede kwaliteit van de spraak (Atal). Men kan het recursieve filter schrijven als:

$$G(z) = \frac{1}{1+a_1z^{-1}+a_2z^{-2}+a_3z^{-3}+\ldots+a_Mz^{-M}} \quad (6)$$

Het nadeel van deze methode voor ons is, zoals in de inleiding werd aangeduid, dat ze niet aansluit bij de formantstructuur die bij fonetici is ingeburgd. Veel spraakperceptie experimenten zijn gedaan met formanten als variabele grootheden. Bovendien is bekend dat een formantbeschrijving van spraak de zuinigste is wat betreft informatie-inhoud. Een derde reden om naar een formantbeschrijving te zoeken is het feit dat er in ons laboratorium een digitale hardware formantensynthetisator aanwezig is, waarmee synthese in real-time mogelijk is. Daardoor krijgt de experimentator een snelle respons van het systeem, hetgeen enorme voordelen heeft.

De overdrachtfunktie van een in serie geschakelde formanten synthetisator kan men schrijven als: (met 5 formanten)

$$G(z) = \frac{1}{(1+p_1z^{-1}+q_1z^{-2})(1+p_2z^{-1}+q_2z^{-2})\ldots} \quad (7)$$

Willen we de spraak m.b.v. formanten resyntiseren, hetgeen we ons ten doel gesteld hebben, dan moeten we het resultaat uit vergelijkingen (6) nog omrekenen naar p,q - data volgens verg (7). Voor dit doel zijn procedures bekend om kwadratische termen van een polynoom af te splitsen.

3. HET ANALYSEPROGRAMMA

Het computerprogramma, dat voor het berekenen van het inverse filter is ontwikkeld, ziet er in grote trekken als volgt uit.

Het spraaksignaal is op de gebruikelijke wijze gedigitaliseerd. (laagdoorlaat-filter, sample and hold-schakelaar en conversie door een 12 bits ADC), en is opgeslagen in het geheugen van de computer. De sample-frequentie bedraagt 10 kHz. Voor een analyseslag worden 250 samples genomen, overeenkomend met 25 ms. Allereerst wordt het aantal nuldoorgangen in het analyse-interval geteld. Dit gegeven wordt gebruikt voor de beslissing of het signaal stemhebbend (periodiek brongeluid gepaard gaand met een gering aantal nuldoorgangen) ofwel stemloos is (ruis als brongeluid en groot aantal nuldoorgangen). Vervolgens wordt het signaal geschaald en dan wordt preemphasis toegepast door een filter:

$$P(z) = 1 - \mu z^{-1}$$

De constante μ wordt voor een analyse interval en is

$$\mu = \frac{R_1}{R_0}$$

Hierin zijn R_1 en R_0 de eerste en nulde autocorrelatiecoëfficienten. Als de spraak stemhebbend is, dan is μ ongeveer 1 en het preemphasis filter is dan een differentiatie. Bij stemloze klanken kan μ gelijk -1 worden en dan is het preemphasis filter gelijk aan een integrator. Het preemphasis filter moet er voor zorgen dat toppen die in het spectrum voorkomen

ongeveer even hoog komen te liggen. Het preemphasis filter is dan ook op te vatten als een invers filter van de eerste orde. Na deze bewerking wordt het stuk spraak vermeningvuldigd met een hammingwindow

$$\chi_n = S_n \ (.54 - .46 \cos \frac{2\pi n}{N-1}) \ , \text{ met } n=1....N.$$

Deze bewerking is gebruikelijk om de ongewenste effecten van een rechthoekig venster te vermijden. Dan volgt het berekenen van het inverse filter met M=10. Daartoe worden 10 autocorrelatie coëfficienten bepaald en wordt het stelsel vergelijkingen (6) opgelost om de filter coëfficienten a_k te vinden. Hiervoor is een snel en recursief algorithme ontwikkeld (MULLER 1973). Zoals in paragraaf 2 is besproken wordt nu dit gevonden filter (van de vormen (6)) in 5 kwadratische termen gesplitst, m.b.v. het Bairstow algorithme. Dit is een iteratieve procedure, die in een enkel geval geen kwadratische term kan vinden. In zo'n geval wordt de term, die bij het vorige analyse interval is gevonden daarvoor in de plaats gezet. Een veel belangrijker probleem is dat de polen niet steeds in een vaste volgorde worden gevonden, terwijl de formanten wel een ordening hebben. Daarom moeten de polen aan de formanten worden toegewezen.

4. HET METEN VAN DE TOONHOOGTE.

Het meten van de toonhoogte is ook een probleem waarvoor sinds decennia naar een oplossing wordt gezocht. Geen van de voorgestelde methoden werkt zonder fouten: de ene methode is zeer gevoelig voor de kwaliteit van het inputsignaal, de andere methode geeft fouten bij kleine signalen en een derde is niet bestand tegen signalen met een sterke tweede harmonische van de grondtoon, enz.

Wij hebben een methode (aangevuld met nog een enkel idee) uit de literatuur (SONDHI 1968) genomen, waarvan bekend is dat hij redelijk betrouwbaar werkt. Voor het meten van de toonhoogte wordt 35 ms (=350 samples) aan het ingangssignaal genomen, om er zeker van te zijn dat minstens twee periodes in het analyse-interval liggen. Dit signaal wordt (niet gewindowed) center-geclipt en vervolgens wordt de autocorrelatiefunctie bepaald. De autocorrelatiefunctie wordt alleen berekend rondom de periode die wordt verwacht. Het maximum in de autocorrelatiefunctie wordt aangewezen als de periode in het ingangssignaal. De grootte van het interval waarin naar het maximum wordt gezocht hangt af van de hoogte van de top in het vorige analyse-interval. Is de top hoog geweest dan is dat een duidelijke periode geweest en dan wordt de breedte van dit venster smal gezet. Deze methode bespaart niet alleen rekentijd maar houdt ook rekening met een zekere continuïteit die in het verloop van de natuurlijke toonhoogte aanwezig is.

5. SLOT OPMERKING

Een korte inleiding werd gegeven van een analyse synthese-systeem gebaseerd op Linear Predictive Coding. Naast de genoemde eigenschap, dat de kwaliteit van de gereproduceerde spraak goed is zijn nog meer eigenschappen interessant die hier onbesproken moesten blijven. We noemen slechts:

- De analyse is bestand tegen storing in het ingangssignaal: het minimaliseringsproces verloopt normaal, zij het met een iets grotere restfout.
- De hier geschetste analysemethode levert een een filterpolynoom die altijd stabiel is.
- Het filter is ook te schrijven als ladderstructuur, die equivalent is aan het mondkanaal opgevat als akoestische buis.

De theorie is in korte tijd (minder dan 10 jaar) gegoeid tot een methode, die op bijna alle gebieden van het spraakonderzoek is doorgedrongen als een gevestigd gereedschap (MARKEL & GRAY 1970).

LITERATUUR

Schroeder, M.R. (1966) Vocoders, Analysis and Synthesis of Speech

Itakura, F. en Saiso, S. (1969) Speech Analysis-Synthesis System based on the partial Autocorrelation Coefficient. Acoust. Soc. of Japan Meeting 1969.

Atal, B.S. en Hanauer, S.L. (1971) Speech Analysis and Synthesis by Linear Predictive of the Speech Wave. J.A.S.A. 50, 1971. 637-655

Sambur, M.R. (1975) An Efficient Linear Prediction Vocoder. BSTJ vol. 54, 1975 , pp. 1693-1723.

Kang, G.S. en Coulter D.C. (1976) 600 BPS Voice Digitizer Int. Conf. ASSP 1976.

Muller, H.F.(1973) Eem methode voor het oplossen van een stelsel vergelijkingen, met een symmetrische coefficienten matrix, IPO Memorandum 122.

Sondhi, M.M. (1968) New Methods of Pitch Extraction IEEE trans. on Audio vol Av-16, 1968, pp. 262-266.

Markel, S.D. en Gray, A.H. jr. (1975) Linear Prediction of Speech,Springer 1976.

Voordracht gehouden op 12 mei 1976 in het Instituut voor Zintuigfysiologie TNO te Soesterberg op een gemeenschappelijke vergadering van het NERG (no. 256), de Benelux-section IEEE en het Nederlands Akoestisch Genootschap.