ANALYSE EN SYNTHESE VAN SPRAAK MIDDELS EEN BANDFILTERSYSTEEM

Ir. Louis C.W. Pols
Instituut voor Zintuigfysiologie TNO, Soesterberg

For the real-time analysis of speech we make use of a parallel set of bandfilters, which outputs are sampled every 10 msec. The raw spectral data are stored on the disk memory of a computer and can be used for subsequent data processing.

However, these data can also be used for a real-time speech synthesis. The synthesis system, also based on a parallel set of bandfilters, gets its data from the computer and all synthesis parameters can be modified under program control.

In this way not just the originally analyzed speech utterances can be regenerated with a good intelligibility, but also any wanted part out of the text can be isolated and/or repeated for careful listening. If wished so, pitch, loudness, or pause duration can be modified.

This system is very appropriate for generating stimuli for speech perception experiments, and is intensively used for studying vowel coarticulation effects.

INLEIDING

In relatie tot het thema van deze serie voordrachten, namelijk de "transmissie en synthese van spraak", wordt bij analyse en synthese van spraak wellicht in eerste instantie gedacht aan de beschrijving van een systeem voor spraaktransmissie, al dan niet met een lage bitrate of bandbreedte. Er zijn echter nog andere redenen waarom men spraak zou willen analyseren en/of synthetiseren.

Men kan spraak *analyseren* om de fundamentele eigenschappen van het spraaksignaal en zijn dynamische variaties te onderzoeken, of om te komen tot automatische spraakherkenning, of spreker herkenning of -verifikatie, of men kan de verkregen informatie in een of andere vorm visueel of taktiel presenteren voor het spraakonderricht aan doven.

Zonder voorafgaande analyse is het toch mogelijk spraak te *synthetiseren* via synthese door regels. Een dergelijk programmeerbaar synthesesysteem kan in het algemeen gebruikt worden voor signaalgeneratie, niet alleen in de vorm van spraak maar eventueel b.v. ook muziek.

Analyse plus synthese van spraak worden gekombineerd in spraaktransmissie systemen, al dan niet inklusief spraakkodering. In het fundamenteel spraakonderzoek kan een dergelijk systeem echter ook met vrucht gebruikt worden voor de (evt. gedeeltelijke) hersynthese van een stuk spraak, waarbij dan al of niet een of meerdere parameters kunnen worden gemodificeerd. Men kan hierbij denken aan intonatie- en koartikulatieonderzoek. Ook kunnen zo stimuli voor perceptieve proeven worden gegenereerd.

Ten behoeve van het experimenteel spraakonderzoek op het I.Z.F. beschikten wij ook graag over een analyse-synthese systeem dat aan de volgende eisen zou moeten voldoen: De synthetische spraak moest van een redelijke kwaliteit zijn, het systeem moest automatisch zijn en onder programmakontrôle kunnen worden gemanipuleerd. Het systeem moest liefst in real time werken en bij voorkeur gebaseerd zijn op een principe representatief voor de spraakperceptie.

Onze keuze is hierbij gevallen op een kanaalvocoder-achtige benadering. Andere mogelijkheden vielen af door moeilijk te objektiveren analysemethodes (formantvocoder), of door de komplexiteit en de te grote computerbehoefte (LPC-vocoders).

In de volgende paragraaf zal het systeem worden beschreven waarna enkele toepassingen van het systeem fragmentarisch zullen worden besproken.

Beschrijving van het analyse-synthese systeem

Zowel aan de analyse- als aan de synthesekant wordt gebruik gemaakt van een parallelle serie bandfilters. In plaats van een transmissieweg fungeert de computer als intermediaire opslag en aktieve datamanipulator. Veel variatie is mogelijk in het type filter, de onderlinge afstand op de frekwentieschaal, de bandbreedte e.d. Wij hebben ons bij onze keuze laten leiden door het beperkte frekwentieoplossend vermogen van het menselijk gehoororgaan, uitgedrukt in de zogenaamde kritische bandbreedte (Plomp, 1976). Tertsfilters hebben een bandbreedte die goed hiermee overeenkomt. Van 400 Hz tot en met 8000 Hz hebben we hier dan ook gebruik van gemaakt. Teneinde de konstante kritische bandbreedte beneden ca. 400 Hz te simuleren zijn nog drie 90-Hz filters met middenfrekwenties van 122, 215 en 307 Hz aan het systeem toegevoegd. Dit

resulteert in 17 filters.

Analyse

Aan de analysekant is aan deze filters nog een breedbandig filter (-3 dB punten bij 32 Hz en 8000 Hz) toegevoegd om het overall nivo van de spraak te kunnen meten. Deze filters worden gevolgd door logarithmische versterkers en omhullende piekdetektoren waarvan de parallelle uitgangen 100 keer per sek. worden bemonsterd met een 20-kanaals multiplexer. Deze informatie gaat via een analoog-digitaal omzetter naar een PDP-15 computer. Voor een blokschema van het analysesysteem zie Fig. 1.

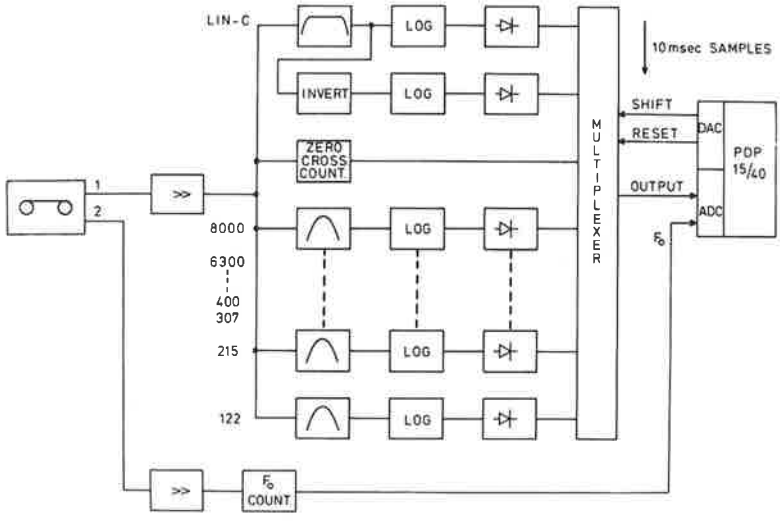

Fig. 1. Blokschema van het analysesysteem.

De ruwe data worden opgeslagen op een schijfgeheugen. Via diverse verwerkings- en displayprogramma's kunnen deze data dan worden bestudeerd. De data kunnen, na eventuele modifikatie, ook gebruikt worden voor een gehele of gedeeltelijke hersynthese. Zowel analyse als synthese gebeuren in real time, de tussenbewerkingen zijn echter off-line. Alvorens het synthesegedeelte te bespreken zullen we enige voorbeelden geven van de verwerkings- en displaymogelijkheden.

Dataverwerking en parameterdisplay

De ruwe gegevens per 10-msek sample kunnen allereerst in de vorm van een getallendisplay worden weergegeven. Het verloop van iedere parameter kan ook als funktie van de tijd zichtbaar gemaakt worden. Fig. 2 is een voorbeeld van een dergelijke display.

Door een pointer over het scherm te bewegen worden de numerieke waarden op ieder moment aangegeven in het rechter gedeelte van de display. Rechtsboven is het bandfilterspektrum op het betreffende tijdmoment zichtbaar. Om het *spektrale* verloop als funktie van de tijd weer te kunnen geven is gebruik gemaakt van een datareduktie techniek. Hiertoe worden de 17 filterwaarden per 10-msek sample beschouwd als de koordinaatwaarden van een punt in een 17-dimensionale ruimte. Ieder bandfilterspektrum is dan een punt in die ruimte, en opeenvolgende spektra, zoals in een woord, vormen een spoor in die ruimte. Met behulp van een principale-komponenten analyse (v.d. Geer, 1967) kan nu een laag-dimensionale subruimte worden gedefinieerd waarin niettemin een zeer groot deel van de variatie in de oorspronkelijke data kan worden beschreven. Voor display doeleinden is een twee-dimensionale weergave uiteraard het meest geeigend.

Fig. 3 geeft ter illustratie hiervan een weergave van de gemiddelde positie in de twee-dimensionale subruimte van een aantal klinkersegmenten geisoleerd uit eenlettergrepige woorden. Als klinkers zijn gekozen de *ie*, *oe*, en *aa*. Een dergelijke weergave vertoont veel verwantschap met die in het zogenaamde formantvlak, is echter ten opzichte van deze eenduidiger en, wat zeer belangrijk is, kan automatisch worden bepaald.

Deze benadering biedt tevens de mogelijkheid tot een visuele representatie van de spektrale informatie ten behoeve van het spraakonderricht aan doven. Een eenvoudiger versie hiervan is in de praktijk met sukses getoetst (Povel, 1974). Vervolgens wil ik in het kort het synthese systeem beschrijven.

Synthese

Uit de literatuur is bekend (Flanagan, 1972) en beide eerste voordrachten hebben dat opnieuw aangetoond (Sluyter; Kuijk en Franssen, dit nummer), dat een voice-excited kanaalvocoder goede kwaliteit spraak kan produceren. Deze benadering was echter voor ons type on-

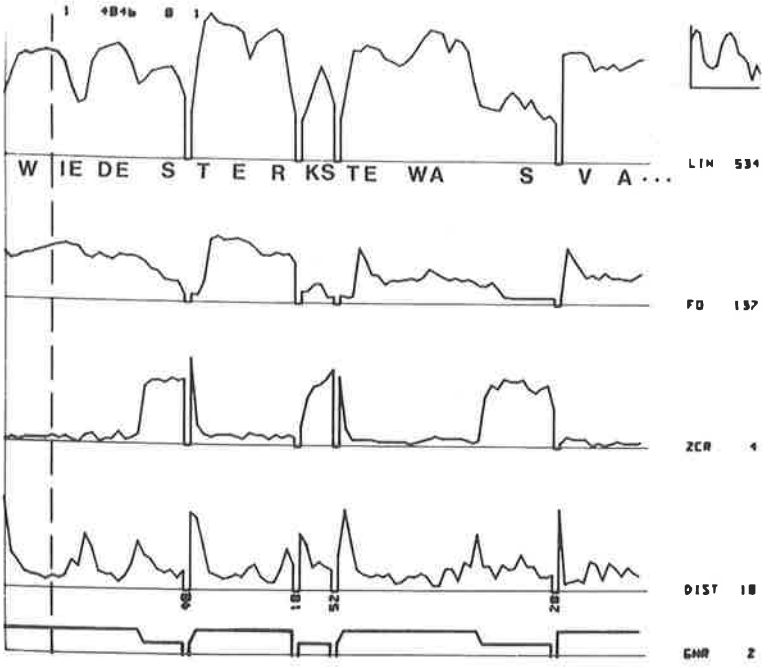

Fig. 2. Weergave van het tijdsafhankelijke verloop van een aantal parameters, zoals gefotografeerd van een computer display. Van boven naar beneden is achtereenvolgens weergegeven: het overall nivo, de grondfrekwentie, het aantal nuldoorgangen, de afstand tussen opeenvolgende spektra in de 17-dimensionale ruimte, en stemhebbend/stemloos. Tevens zijn op de betreffende plaatsen de pauzeduren in msek aangegeven. De stippellijn is een verplaatsbare pointer. In de rechterhelft van de display zijn de met dat tijdmoment overeenkomstige numerieke waarden van de diverse parameters aangegeven. Tenslotte is rechts bovenin het momentane bandfilterspektrum weergegeven. Het stuk tekst waarvan hier de analyseparameters zichtbaar zijn is: "Wie de sterkste was van ... ".

derzoek, waarbij we ook de toonhoogte vrijelijk wilden kunnen modificeren, minder geschikt. Wij werken dan ook met een vaste periodieke of ruisbron, voor het genereren van de stemhebbende en stemloze spraakklanken. Echter in plaats van de uitgangen van de filters te moduleren, introduceren wij de spektrale variatie door het bronsignaal zelf per filter te varieren. De synthetisator is een perifeer apparaat dat door middel van het op ons instituut ontwikkelde digitale input-output systeem vanuit de computer per 10 msek van nieuwe data wordt voorzien, waarna het apparaat zelfstandig de signaalgeneratie verzorgt.

De voor de synthese benodigde toonhoogte informatie wordt tijdens de analyse verkregen middels een keelmikrofoonsignaal. Door onder programmakontrôle de parameters nodig voor de synthese te modificeren kan de teruggegenereerde spraak op vele manieren worden beïnvloed. Zo kan een klein gedeelte worden geisoleerd en in detail uitgeluisterd, door herhaald genereren en/of uitrekken van de tijdschaal, zonder spektrale vervorming. De binnen- en tussenwoord pauzes kunnen worden benadrukt door ze te verlengen, of juist geheel over te slaan. De intonatie van de zin kan worden gewijzigd door modifikaties in het verloop van de toonhoogte aan te brengen.

Ook kan rekenkundig of eventueel in hardware een datareduktie op de basisgegevens worden toegepast, waarna hersynthese met een lagere bitrate kan worden gerealiseer. Tijdens de voordracht op 12 mei 1976 werden door middel van een bandopname een aantal van deze mogelijkheden gedemonstreerd.

Alhoewel flexibiliteit en eenvoudige modificeerbaarheid eerste vereisten van het systeem zijn, willen we toch ook iets zeggen over de spraakverstaanbaarheid van teruggegenereerde spraak. De spraakverstaanbaarheid is op de beproefde wijze gemeten via lijsten met eenlettergrepige nonsens woorden (fonetisch gebalanceerd = PB words), ingesproken door vijf verschillende sprekers, en beluisterd door vijf verschillende luisteraars. De gemiddelde PB-word verstaanbaarheid bedraagt 77,8%. Dit is voor onze toepassingen ruim voldoende en garandeert een voortreffelijke zinsverstaanbaarheid.

Wanneer datareduktie wordt toegepast neemt de woordverstaanbaarheid uiteraard af, en wordt 44,1% wanneer 3-dimensionale in plaats van de oorspronkelijke 17-dimensionale spektrale informatie wordt gebruikt.

Ons systeem biedt ook de mogelijkheid tot een zogenaamde pattern matching vocoder (Flanagan, 1972). Hierbij wordt gewerkt met een beperkt aantal diskrete spektra die gelabeled zijn, alleen deze nummers worden overgestuurd. In plaats van met diskrete spektra zou in ons geval met diskrete posities in een laag-dimensionale subruimte worden volstaan.

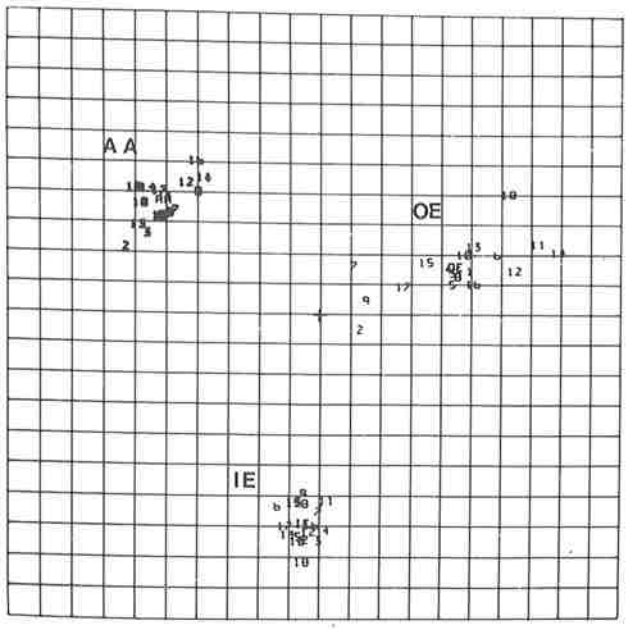

Fig. 3. Computer display van de gemiddelde posities van 18 klinkersegmenten geisoleerd uit eenlettergrepige woorden voor de klinkers *ie*, *oe*, en *aa*, weergegeven in de twee-dimensionale spektrale subruimte.

Tot slot een summiere beschrijving van twee typen onderzoek waarvoor het analyse-synthese systeem wordt gebruikt.

Onderzoek met het analyse-synthese systeem.

Teneinde enig inzicht te krijgen in de belangrijkheid van bepaalde frekwentiegebieden voor de perceptie van bepaalde klinkers is een benoemingseksperiment uitgevoerd met selektief gemaskeerde klinkers. Met behulp van de synthetisator werden hiertoe korte klinkers gegenereerd waarbij de spektrale informatie in één van de filters was vervangen door een band ruis. Als dit gebeurt in een frekwentiegebied dat niet erg bijdraagt tot de korrekte identificeerbaarheid van een bepaalde klinker, dan zal de maskering nauwelijks tot foutieve benoemingen leiden. Wanneer echter essentiele spektrale informatie wordt gemaskeerd dan zal dit leiden tot specifieke foutieve benoemingen (Pols, 1975).

Een heel ander type onderzoek betreft de fysische en perceptieve verschillen tussen klinkers in eenlettergrepige woorden. De variabelen die hierbij een rol spelen zijn de verschillnede Nederlandse klinkers, de verschillende sprekers, en de verschillende medeklinkeromgevingen waarin de klinkers kunnen worden geplaatst.

Het analyse-synthese systeem wordt gebruikt voor het analyseren en isoleren van de klinkersegmenten, en voor het hersynthetiseren van deze segmenten voor benoemingsproeven. Niet alleen de tweeklanken *ei*, *au* en *ui* blijken een sterk dynamisch verloop te hebben, maar ook klinkers als de *ee*, *oo* en *eu*, waarvan het verloop zich dan ook nog sterk wijzigt als deze klinkers voorafgaan aan de *r*.

Soesterberg, mei 1976

LITERATUUR

Flanagan, J.L. (1972). *Speech analysis synthesis and perception*, Springer Verlag, Berlin, 2nd edition.

v.d. Geer, J.P. (1967). *Inleiding in de multivariate analyse*, Van Loghum Slaterus, Arnhem.

Kuijk, K.E. en Fransen, N.V. "Reductie van spraakbandbreedte m.b.v. een vocoder, welke gebruik maakt van vooruitregeling", dit nummer.

Plomp, R. (1976). "Auditieve functies", Hoofdstuk 7,2 in *Handboek der Psychonomie*, Eds. Michon, J.A., Eijkman, F.G.J. en de Klerk, L.F.W., Van Loghum Slaterus, Deventer.

Pols, L.C.W. (1975). "Dominant spectral regions in vowel perception", Paper 347 of the 8th Congress of Phonetic Sciences, Leeds.

Povel, D.J.L. (1974). "Articulation correction of the deaf by means of visually displayed acoustic information", dissertatie K.U. Nijmegen.

Sluyter, R.J. (1976). "Een eenvoudige vocoder voor digitale spraakoverdracht", dit nummer.

Voordracht gehouden op 12 mei 1976 tijdens een gezamenlijke vergadering van het NERG (256ste werkvergadering), de Benelux-section IEEE, en het NAG, op het Instituut voor Zintuigfysiologie TNO, te Soesterberg.