Spraaksynthese : stand van zaken en toekomst

SPRAAKSYNTHESE: STAND VAN ZAKEN EN TOEKOMST

Ir. L.F. Willems
Instituut voor Perceptie Onderzoek

The state of the art of speech synthesis is described in this paper. The application of speech synthesis in speaking machines is coming nearer through the availability of speech synthesis chips. The text to speech conversion problem is, however, not yet solved satisfactorily.

1. INLEIDING

Spraak is voor ons mensen een heel natuurlijk communicatiemiddel. Wij maken er veelvuldig gebruik van en het is ook te verwachten, dat bij geavanceerde en mens-vriendelijke communicatie tussen mens en machine spraak een grote rol zal spelen. Spraaksynthese, het opwekken van kunstmatige spraakklanken, heeft hier het doel om boodschappen vanuit een apparaat te produceren, zodat de menselijke gebruiker ze kan verstaan en erop kan reageren.

De mens heeft al sinds lang de spraak bestudeerd en ook geprobeerd spraakklanken na te bootsen. Een van de eersten was Wolfgang von Kempelen, die in 1791 een spreekmachine construeerde, waarmee hij, zoals hij schreef:

> '...alle Latijnse, Franse en Italiaanse woorden zonder uitzondering kon namaken...zoals bijv. Papa, Maman, Marianna, Maladie, enz. ...'

Von Kempelen had voor de bediening van zijn mechanische spreekmachine beide handen en de nodige vingervlugheid nodig om dit te kunnen presteren. Na de uitvinding van de telefoon en toen deze ingevoerd raakte ontstond belangstelling voor het (electrische) spraaksignaal van de kant van de telefooningenieurs.
In de dertiger jaren heeft Homer Dudley van de Bell Labs pionierswerk verricht. Hij maakte de Vocoder en de Voder. Rond die tijd werd de geluidsspectrograaf ontwikkeld, waarmee het veranderende spectrum als functie van de tijd zichtbaar kon worden gemaakt.
Na de tweede wereldoorlog was er op vele gebieden van de wetenschap een opleving, óók op het gebied van het spraakonderzoek. Er was toen grote belangstelling voor spraakproduktie (articulatie, akoestiek van het mondkanaal, synthetische spraak) en ook voor de waarneming van spraak door de mens (auditieve filtering, Motor Theory of Speech Perception, enz.). In het begin van de 70-er jaren is de LPC-techniek ontwikkeld en nu beleven we de tijd van de stormachtige ontwikkeling van de electronica, waardoor fantastische mogelijkheden beschikbaar komen (computing power, (V)LSI-schakelingen, etc.).

We zullen in dit artikel allereerst ingaan op een aantal doorsnijdingen die men kan maken in het gebied van de spraaksynthese. De eerste doorsnijding heeft te maken met de techniek:

- golfvormcodering, versus:
- resynthese van geanalyseerde spraak, versus:
- spraaksynthese door regels.

Een tweede doorsnijding heeft te maken met toepassingsgebieden:

- vaste boodschappen
- variabele boodschappen
- willekeurige tekst uitspreken

Een derde doorsnijding is:

- complexiteit
- benodigde bitrate of geheugencapaciteit
- spraakkwaliteit

Vervolgens willen we nagaan hoe spraakklanken door de mens gemaakt worden om daaruit mogelijkerwijs inspiratie op te doen voor de manier waarop we spraakklanken kunnen nabootsen.
Het zwaartepunt zal vervolgens liggen bij de middelen om spraak te resynthetiseren en de mogelijkheid voor spraaksynthese door regels.

2. ENKELE ALGEMENE OPMERKINGEN OVER SPRAAKSYNTHESE

Voordat we zijn ingegaan op de methoden om spraak te synthetiseren willen we enkele algemene opmerkingen maken die de verschillende mogelijkheden en aspecten in hun samenhang tonen.
Om boodschappen vanuit een apparaat ten gehore te brengen hebben we nodig: een geheugen in welke vorm dan ook en een omzetter om de gecodeerde spraakgegevens die in het geheugen zijn opgeslagen weer in hoorbare signalen terug te brengen. Over het geheugen zullen we niet veel zeggen: het zou een tape kunnen zijn, maar meestal is het een digitaal geheugen (ROM, RAM, floppy disk, enz.). De toe te passen omzetters kunnen we globaal in een drietal groepen onderverdelen:

a. Golfvormcodering. Deze kunnen boodschappen reproduceren die van tevoren zijn opgenomen en gecodeerd en waarvan de golfvorm volgens een of ander recept is beschreven. Dat kan zijn PCM, waarvoor toch zo'n 64 kbit/sec. nodig is, tot aan de andere kant van de schaal LPC met multipuls-excitatie waarbij met 9600 bit/sec. al zeer goede spraakkwaliteit kan worden bereikt. Bij deze manier van opslaan van de spraak, is

het achteraf, bij het ten gehore brengen ervan, niet meer mogelijk wijzigingen in de boodschap aan te brengen (Voor de verschillende methodes van spraakcodering zie Deprettere, deze uitgave).

b. Resynthese van geanalyseerde spraak. Hierbij worden spraakboodschappen van tevoren opgenomen en geanalyseerd om er een parametrische beschrijving van te maken. Bij het ten gehore brengen van de zo opgeslagen spraakboodschappen moet men de klanken weer op grond van die parametrische beschrijving 'terug opbouwen' (= resynthetiseren). Het voordeel van deze methode is dat naast een aanzienlijke reductie van de benodigde bitrate bij de resynthese de spraakklanken nog gewijzigd kunnen worden (door namelijk vóór resynthese een of meerdere van die parameters te wijzigen). Dit is van groot belang om woorden of andere gebruikte fragmenten aan te passen aan de omgeving van de zin waarin ze zijn geplaatst. Dat geldt voor de duur van de klanken en vooral ook voor de toonhoogte. Het is gemakkelijk aan te tonen dat een dergelijke aanpassing de natuurlijkheid van de geproduceerde spraak aanzienlijk kan verbeteren.
We zullen de resynthese van spraak uitvoerig behandelen in paragraaf 5.

c. Spraaksynthese door regels. Bij deze vorm van spraaksynthese gaat men niet uit van van tevoren opgenomen spraak, maar de spraakboodschap wordt op basis van de tekst of fonetische tekst volledig kunstmatig gemaakt. Meestal gebeurt dat door een gering aantal kleine eenheden achter elkaar te schakelen. Men moet dan regels hanteren om de overgangen van de gebruikte eenheden op de juiste wijze te laten verlopen, om de gemaakte spraak zo natuurlijk mogelijk te laten klinken. Daarnaast moet ook de bovengenoemde aanpassing van de duuropbouw en het verloop van de toonhoogte plaatsvinden. Ook op deze methode van spraaksynthese door regels zullen we nader ingaan in paragraaf 7.

Een tweede doorsnijding van het spraaksynthesegebied heeft te maken met de toepassingen. Er is nogal wat verschil tussen een sprekende thermometer en sprekende telefoongids wat betreft het te kiezen systeem, de benodigde geheugenruimte enz.

a. Vaste boodschappen. Er zijn een aantal toepassingen waarin men gebruik maakt van een beperkt aantal vaste boodschappen. Enkele voorbeelden hiervan zijn: waarschuwingen in de auto, zoals 'Opgelet! Uw Oliepeil is te laag. Ga onmiddellijk naar een garage', of de bovengenoemde sprekende thermometer voor een blinde: 'Het is', 'negentien' 'graden'.

b. Variabele boodschappen. Er zijn toepassingen waarin de te geven boodschappen kunnen worden samengesteld uit korte fragmenten zoals woorden en woordgroepen, maar waarbij de fragmenten nog moeten worden aangepast aan de omgeving waarin ze voorkomen. Voorbeelden zijn: een sprekende klok die zegt: 'Het is nu' 'twintig' 'uur' 'dertien' of die kan zeggen 'Het is nu' 'dertien' 'uur' 'zeven'. In deze twee zinnen zal het woord 'dertien' verschillend klinken afhankelijk van de plaats in de zin. Een ander voorbeeld is het gesproken weerbericht of weerpraatje. Deze kunnen worden samengesteld met een betrekkelijk gering aantal woorden, echter ook hier is het nodig de woorden aan te passen aan de plaats in de zin, plaats van de klemtoon, enz.

c. Willekeurige tekst uitspreken. Dit zijn toepassingen waarin men geen van tevoren opgenomen spraak kan gebruiken, omdat óf de geheugenruimte niet toereikend is (sprekende encyclopedie, sprekend telefoonboek) óf omdat de spraakboodschappen nog niet vastliggen (geavanceerde informatiedialogen, spreekhulpmiddelen voor spraakgestoorden).

Tenslotte zijn er nog een drietal grootheden, die onderling afhankelijk zijn en die een belangrijke rol spelen bij de keuze van de een of andere oplossing voor een bepaald spraakoutputprobleem. Deze zijn: de complexiteit, de benodigde bitrate en de spraakkwaliteit.

a. De complexiteit van een codeer- of syntheseschakeling bepaalt vaak de prijs van het uiteindelijke apparaat, maar hangt nauw samen met de benodigde bitrate en dus ook met de grootte van het geheugen.

b. De benodigde bitrate hangt op zijn beurt weer heel sterk samen met de bereikte kwaliteit van de geproduceerde spraak. De uiterste grenzen waarbinnen de bitrate zal liggen zijn: aan de hoge kant ongeveer 100 kbit/sec. (of meer) en aan de lage kant ca 100 bit/sec. (Deze lage grens kan men afschatten door te bedenken dat er 40 verschillende spraakklanken zijn en dat per seconde zo'n 10 à 15 verschillende klanken door een spreker worden gezegd. Dan komt men ongeveer tot 100 bit/sec. informatie).

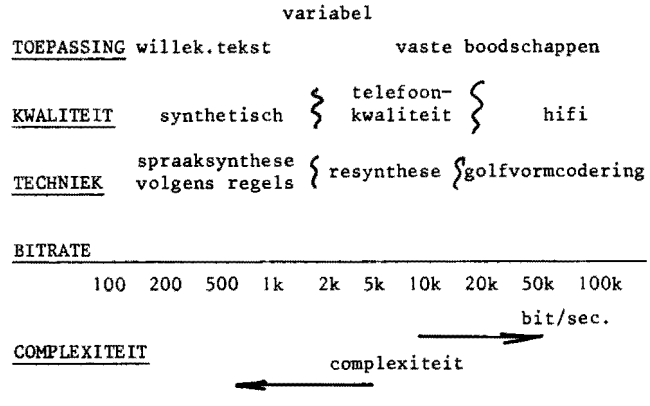

Fig. 1: Overzicht van verschillende grootheden, die in paragraaf 2 zijn besproken.

c. De spraakkwaliteit is natuurlijk een belangrijke eigenschap van een systeem. Er zijn geen objectieve methoden om de spraakkwaliteit te meten. Door middel van meestal tijdrovende luisterproeven kan men spraakkwaliteit kwantificeren (Steeneken, deze uitgave). Overigens is in de loop der jaren de spraakkwaliteit bij een bepaalde bitrate steeds toegenomen. De vooruitgang op dit gebied komt dus tot uitdrukking in óf een lagere bitrate óf een hogere spraakkwaliteit.

In Fig. 1 is getracht de hier genoemde aspecten in beeld te brengen.

3. NATUURLIJKE SPRAAK

Men kan zeggen dat het spraakgeluid wordt gevormd door een veranderlijke geluidsbron en een veranderlijk akoestisch filter dat het brongeluid wijzigt.
Voor de stemhebbende klanken (klinkers en een aantal medeklinkers als: m, n, l, b, d) ontstaat het brongeluid doordat de stembanden trillen. Deze trilling wordt veroorzaakt door een luchtdruk in de longen, die de stembanden uit elkaar duwt; dan gaat er lucht stromen; hierdoor ontstaat t.g.v. het Bernouilli-effect tussen de stembanden een onderdruk, waardoor de stembanden weer dichtgaan, daarbij ook nog geholpen door veerkracht in de stembanden. Hierdoor ontstaan luchtdrukimpulsen met een zekere herhalingsfrequentie. De bronfrequentie bepaalt de waargenomen toonhoogte. Een spreker regelt de bronfrequentie en dus de toonhoogte d.m.v. de mechanische spanning in de stembanden. De luidheid van de spraak wordt voornamelijk bepaald door het luchtdrukverschil tussen onder en boven de stembanden. Het filter voor de stemhebbende klanken is de mond- en keelholte. Als nasale klanken (m en n) worden gemaakt bestaat het filter ook nog uit de neusholte, omdat het zachte verhemelte het neuskanaal opent. Tijdens het spreken verandert voortdurend het mondkanaal van vorm, door bewegingen van de tong, kaak, enz. en dus verandert de filterwerking en daarom ook de klankkleur van het spraakgeluid.

Voor de stemloze wrijfklanken (f, s en g) is het brongeluid ruis die ontstaat door turbulentie van de luchtstroom uit de longen door een vernauwing in het mondkanaal. Voor de v en de z zijn er twee geluidsbronnen: trillende stembanden en luchtturbulentie. Het akoestisch filter bij deze klanken wordt gevormd door de holtes vóór en achter de vernauwing.
Bij plofklanken wordt het mondkanaal gedurende 50 ms tot 100 ms volledig afgesloten en dan weer geopend. Door de plotseling vrijkomende lucht wordt gedurende een korte tijd een ruisgeluid gevormd. In tegenstelling tot deze stemloze plofklanken (p, b, k) blijven bij de stemhebbende plofklanken (b, d) tijdens de afsluiting de stembanden juist doortrillen. Het akoestisch filter bij plofklanken wordt gevormd door de holtes vóór en achter de afsluiting.

4. SYNTHETISCHE SPRAAKKLANKEN

Bij het nabootsen van spraakklanken kan men ook een geluidsbron gevolgd door een filter nemen om zo spraakgeluid te vormen. In dit bron-filter-model wordt de bron U gevolgd door twee filters: het filter O gevormd door de keel- en mondholte en het filter R, dat de straling van het geluid bij de mondopening beschrijft (zie Fig. 2).

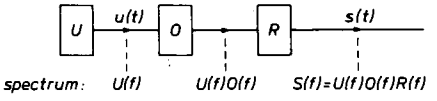

Fig. 2: Blokdiagram van het bron-filter-model.

Het brongeluid U is ofwel een reeks pulsen met een zekere herhalingsfrequentie ofwel ruis.
De overdrachtsfunctie O(f) is voornamelijk verantwoordelijk voor de klankkleur van het geluid. De mondkeelholte is te beschouwen als een wat grillig gevormde buis, die aan een kant -bij de stembanden- vrijwel gesloten is en aan de andere kant open. De overdrachtsfunctie van een dergelijke buis vertoont pieken bij de resonantiefrequenties. Deze pieken noemt men formanten. Elke formant wordt gekarakteriseerd door een middenfrequentie en een bandbreedte. Voor de waarneming van spraak zijn in het algemeen niet meer dan vijf formanten in het gebied tussen 0 Hz en 5000 Hz van belang. Deze worden over het algemeen aangeduid met F1 t/m F5.

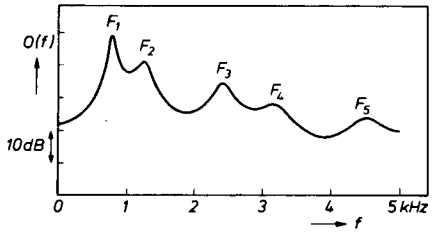

Fig. 3: Overdrachtsfunctie O(f) van een bepaalde mondstand met de formanten F1 t/m F5.

Apparaten of algorithmen voor spraaksynthese kan men baseren op dit bron-filter-model (zie Fig. 4). Als brongeluid neemt men ofwel periodieke impulsen met een zekere herhalingsfrequentie ofwel witte ruis. Dit brongeluid krijgt de gewenste sterkte door volume-instelling en wordt vervolgens gefilterd door een filter O'(f). In de overdrachtskarakteristiek van O' zijn verdisconteerd de veranderlijke eigenschappen van de mondkeelholte en verder de constante eigenschappen van de straling bij de mondopening (R in Fig. 2) en constante spectrale eigenschappen van de geluidsbron.

Voor stemhebbende signalen zijn in Fig. 4 enkele signalen met bijbehorende spectra geschetst.

Men zal bij het proces van spreken de mondstand steeds veranderen en dus zal ook het synthesemodel voortdurend veranderende parameters krijgen toegestuurd die het brongeluid en de overdrachtskarakteristiek bepalen. De snelheid waarmee de articulatoren bewegen is beperkt en dus kan men ook de sturende grootheden voor het synthesemodel ook met een overeenkomstig langzame snelheid veranderen. Dit is dan ook de reden waarom men een dergelijke parametrische beschrijving van het spraaksignaal met een geringere informatiestroom kan beschrijven dan het microfoonsignaal.

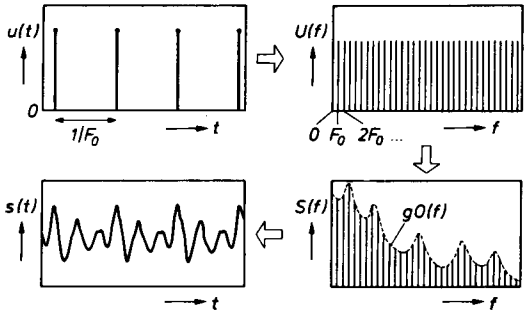

Fig. 4: Signalen en spectra in het synthesemodel voor stemhebbende klanken. Het brongeluid is u(t): periodieke deltapulsen met herhalingsfrequentie F_0. Het spectrum U(f) krijgt door het filter O'(f) de juiste spectrale samenstelling. Tenslotte is S(t) het gemaakte spraaksignaal.

5. SPRAAKRESYNTHESE

Het is mogelijk om de sturende grootheden voor zo'n synthesemodel uit natuurlijke spraak te bepalen. Op de analysemethoden zullen wij hier niet ingaan. In Fig. 5 is een compleet analyseresultaat getekend voor een Nederlandse zin gesproken door een mannenstem.
De analyse wordt 100 keer per seconde uitgevoerd, zodat een analyseresultaat beschikbaar is voor elke 10 ms. Deze frequentie voor het herhalen van de analyse is gebleken voldoende te zijn om het veranderende spraaksignaal te bemonsteren. De analyse wordt uitgevoerd over een spraaksegment van ongeveer 30 ms. In de bovenste twee hokken in Fig. 5 zijn de gegevens voor de geluidsbron weergegeven. De sterkte van het geluid G en de herhalingsfrequentie F0 van de stemhebbende geluidsbron. Tussen de hokken is nog aangegeven wanneer de ruisbron moet worden gebruikt.

In de onderste rechthoek zijn de gegevens geschetst die nodig zijn om het variabele filter in te stellen. Voor elk tijdstip (van 10 ms) worden de middenfrequenties van 5 formanten gegeven met bijbehorende kwaliteitsfactor. Met behulp van deze parametrische beschrijving is

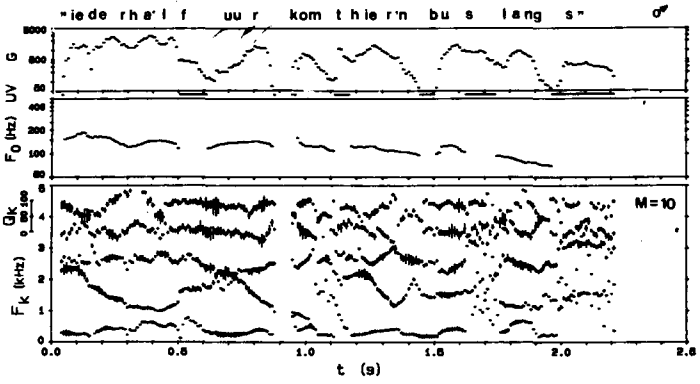

Fig. 5: Compleet analyseresultaat voor een mannenstem. Verklaring in de tekst.

het mogelijk heel behoorlijk spraak te resynthetiseren. Ook is het mogelijk om de parametrische beschrijving voor resynthese te wijzigen, bijvoorbeeld wat betreft de F0 (verantwoordelijk voor de waargenomen toonhoogte) en wat betreft de duur van spraaksegmenten. Dit was immers van groot belang om de geresynthetiseerde boodschappen natuurlijk te laten klinken.

Het variabel filter in het synthesemodel kan op verschillende wijzen geïmplementeerd worden: bijv. als ladderfilter of als spectrumshaper met bandfilters en amplituderegeling voor elk kanaal (zoals in kanaalvocoders). Het is echter bekend dat de hier gebruikte codering m.b.v. formanten de zuinigste beschrijving is. Een nadeel is echter dat de bepaling van de formanten uit natuurlijke spraak niet zonder problemen is.

Het verlies aan spraakkwaliteit dat men kan beluisteren bij deze spraakresynthese is te wijten aan het feit dat het bron-filter-model niet in staat is om de akoestische verschijnselen van het proces van spreken voldoende nauwkeurig te beschrijven. Zo zal het functioneren van de stembanden niet onafhankelijk zijn van de mondkeelholte. Ook is het gebruikte filter met een aantal resonantiepieken niet in staat de akoestische invloed van het neuskanaal te beschrijven of de invloed van de holtes achter de afsluiting bij wrijfklanken. Ook bij de bepaling van de verschillende grootheden gaat men ervan uit dat gedurende het analyse-interval (ca 30 ms) het signaal stationair is. Deze aanname zal zeker niet gelden bij plofklanken en andere snelle veranderingen.

6. SPRAAKCHIPS

Als men zo'n parametrische beschrijving heeft gemaakt, kan men met luisterexperimenten nagaan of op de codering van de gegevens kan worden bezuinigd. Eerst door de nauwkeurigheid waarmee elk gegeven wordt vastgelegd te beperken en ten tweede door de frequentie te beperken waarmee

de gegevens door nieuwe worden vervangen. Men kan nog verstaanbare spraak resynthetiseren met een bitrate van ongeveer 1000 bits/sec.

De laatste jaren hebben verschillende fabrikanten spraaksynthesechips gemaakt en op de markt gebracht, waarop een complete spraaksyntheseschakeling, meestal in digitale techniek, is ondergebracht. Ik zal hier een spraakchip: de MEA8000 van Philips, nader beschrijven die gebaseerd is op de al eerder beschreven codering in formanten. Het blokschema van de MEA8000 is weergegeven in Fig. 6. De codering voor deze chip is weergegeven in de onderstaande tabel I.

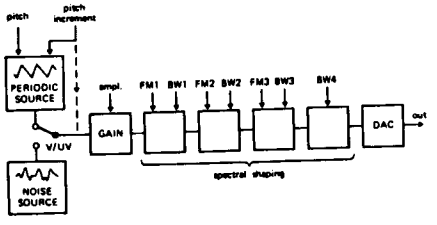

Fig. 6: Blokschema van de MEA8000 spraaksynthesechip.

Tabel I.

Afkorting	bits	parameter
FD	2	spraakframe duur (8, 16, 32, 64 ms)
AM	4	amplitude in log eenheden
PI	5	toename toonhoogte en ruis-keuze
F1	5	frequentie van formant 1
F2	5	frequentie van formant 2
F3	3	frequentie van formant 3
B1	2	bandbreedte van formant 1
B2	2	bandbreedte van formant 2
B3	2	bandbreedte van formant 3
B4	2	bandbreedte van formant 4
Totaal	32	

De frequentie van de vierde formant is vastgelegd op 3500 Hz. De frameduur wordt ook gecodeerd en met 2 bits kan men kiezen tussen 8 ms, 16 ms, 32 ms en 64 ms. Hieruit volgt dat de hoogste bitrate welke aan deze chip kan worden toegevoerd 4000 bits/sec. is (alle frameduren 8 ms) en de laagste bitrate is 500 bits/sec. (alle frameduren 64 ms). Het is de bedoeling om de frameduur aan te passen aan de mate waarmee het spraaksignaal zelf verandert: bij een snelle overgang gebruike men korte segmenten en in stabiele stukken gebruike men lange segmenten). In de chip worden de grootheden 8 keer per frame geïnterpoleerd om zodoende grote overgangen (die zeker bij lange frameduren zouden optreden) glad te strijken. In de praktijk liggen de benodigde bitrates voor goed verstaanbare spraak tussen de 1000 en 2000 bits/sec. In een toepassing van zo'n spraakchip heeft men naast deze chip ook nog nodig een geheugen (PROM of ROM) waarin de gecodeerde spraak ligt opgeslagen en een microprocessor die het datatransport regelt. Voor een toepassing zal men een aantal boodschappen of fragmenten van meldingen (denk aan een sprekende klok) van tevoren door een spreker laten zeggen, laten analyseren door een computer of spraakontwikkelsysteem (kan door de fabrikant van de chip worden gedaan) en tenslotte in een geheugen laten vastleggen. Er zijn intussen een groot aantal van dergelijke spraaksynthesechips te koop. De toepassing ervan komt echter traag op gang.

7. WILLEKEURIGE TEKST UITSPREKEN

Wil men willekeurige teksten laten uitspreken door een automaat, dan moet de tekst eerst omgezet worden in een fonetische transcriptie om vervolgens door een spraaksynthese-door-regels-systeem te worden omgezet in verstaanbare spraak. Het eerste probleem: de omzetting van tekst in een fonetische transcriptie beschouw ik hier als gegeven (zie Boot, deze uitgave). Ik ga ook ervan uit dat de fonetische transcriptie is voorzien van indicaties waar lettergrepen klemtoon krijgen.
Bij het tweede probleem, dat van het spraaksynthese-door-regels-systeem staat centraal de vraag uit welke eenheden zal men de spraakuiting samenstellen. Neemt men weinig eenheden, zoals de elementaire spraakklanken (soms fonemen genoemd) dan heeft men er slechts weinig nodig (ca 40), maar de regels die nodig zijn om vervolgens de klanken aan te passen aan hun omgeving zullen nogal ingewikkeld zijn. Vooral de overgang van de ene klank naar de andere is moeilijk met behulp van regels te beschrijven. Neemt men daarentegen grote eenheden bijv. woorden dan is het duidelijk dat men zeer veel geheugenruimte nodig heeft voor de opslag, maar dat de regels voor aanpassing aan de omgeving veel simpeler zullen zijn.

Een aardig compromis, dat de laatste jaren nogal wat aandacht krijgt, lijkt te zijn difoon-synthese. De eenheden zijn difonen: stukje klank + overgang + stukje volgende klank. Daardoor heeft men de overgangen niet door regels hoeven te beschrijven en het bovengenoemde probleem is zodoende omzeild. Het aan elkaar koppelen van spraaksegmenten in de meer stabiele stukken geeft vrijwel geen problemen. Voor een dergelijk systeem heeft men ca 1600 difonen nodig.
In een systeem dat door ons gebouwd wordt, waarvoor de input is: fonetische tekst met klemtoontekens en de genoemde spraakchip het uitvoerorgaan is, wordt ongeveer 50 kbyte gebruikt voor de opslag van de difonen. De codering van de spraakgegevens voor de difonen is dezelfde als in paragraaf 5 is beschreven. In Fig. 7 is de codering geschetst van het woord 'banaan', samengesteld uit difonen. Op de difoongrenzen, waar de fragmenten aan el-

kaar gekoppeld zijn, kan men kleine discontinuïteiten zien, maar men kan ze vrijwel niet horen.

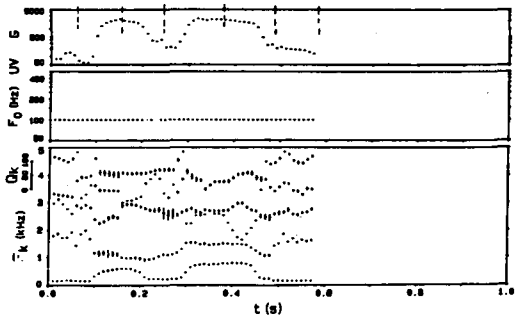

Fig. 7: Parametrische beschrijving voor het woord 'banaan', verkregen door difoonconcatenatie. De difoongrenzen zijn aangegeven met stippellijnen.

De gegevens voor de difonen zijn gehaald uit beklemtoond uitgesproken lettergrepen uit onzinwoorden als 'nenaane'. Hieruit kan men het difoon 'naa' en het difoon 'aan' halen. Heeft men nu een zin samengesteld uit dergelijke difoonfragmenten dan klinkt zoiets nog helemaal monotoon. Een grote sprong in natuurlijkheid krijgt men door de toonhoogte aan te passen aan de intonatie van een dergelijke Nederlandse zin. Ook zal aanpassing van de duren van de segmenten aan de plaats in de zin verbetering geven. Immers de difonen zijn allemaal gehaald uit beklemtoonde lettergrepen en ze komen in een zin ook voor op niet beklemtoonde posities.

8. SLOTOPMERKINGEN

Het spraakonderzoek krijgt tegenwoordig nogal wat aandacht. Dit zal onder andere ertoe leiden dat de kwaliteit van synthetische spraak steeds zal verbeteren. Ik wil hier enkele mogelijkheden noemen, die er zijn om het proces van spreken nauwkeuriger in kaart te brengen en zodoende de kunst van het opwekken van synthetische spraak vooruit te helpen.

- Verlaten van bron-filter-model. De generator van het brongeluid (de stemband-oscillator) wordt onafhankelijk beschouwd van het akoestische filter (het mondkanaal). De aannames die hierin worden gemaakt vormen een te grote beperking. Ingewikkelder modellen betekenen echter ook complexere synthesetechnieken en moeizamere analysemethoden.
- De aanname van de (quasi-)stationariteit vormt ook een grote beperking. Er zijn te veel spraaksegmenten, die hierdoor niet of slecht worden weergegeven in de analyseresultaten.
- Er is nog betrekkelijk weinig kennis omtrent de juiste duuropbouw van spraakuitingen. Dit komt o.a. tot uiting in de difoonconcatenatie.
- Een groot probleem, dat wel aandacht begint te krijgen, maar toch nog niet opgelost is, is de fonetische transcriptie of anders gezegd de grafeem-foneem-omzetter.

De vele aandacht voor spraak zal ook tot uiting komen in meer toepassingen dan tot nu toe zijn gemaakt. Bekend zijn: 'Speak and Spell' van Texas Instruments dat een zekere pioniersrol heeft vervuld en voorts het sprekende dashboard van een type Renault.
Dat er nog ruimte voor eenvoudige toepassingen is blijkt wel uit het feit dat bij de landing van de eerste space shuttle de ene hooggetrainde piloot aan de andere piloot de stand van de hoogtemeter moest voorlezen.

Ik ben van mening dat de toepassingen van spraaksynthese pas goed op gang zullen komen, als de apparaten ook onze spraak kunnen verstaan, zodat er een natuurlijke dialoog mogelijk is tussen de mens en de machine.

Tijdens de voordracht werd een en ander met geluidsvoorbeelden geïllustreerd.

Voor verdere lezing aanbevolen:

Flanagan, J.L. and Rabiner, L.R. (eds). Speech synthesis. Benchmark Paper in Acoustics.

Hart, J. 't et al. Manipulaties met spraakgeluid. Philips Technisch Tijdschrift 40, no. 4, 108-119.

MEA8000 voice synthesizer: principles and interfacing. Techn. Publication 101, Elcoma.

Witten, Ian H. Principles o- computer speech. 1982. Academic Press.

Enkele artikelen in Databus n2 7/8, juli/augustus 1982.

Voordracht gehouden tijdens de 319e werkvergadering.