Sprekend jezelf. Een verslag van twee onderzoeken naar sprekerkenmerken.

Henk van den Heuvel en Hans Kraayeveld

In this paper we summarise two recent dissertations by the authors concerning speaker characteristics on the acoustical-phonetic level of the speech signal. Van den Heuvel (1996) deals with speaker characteristics in the realisations of individual phonemes, whereas Kraayeveld (1997) concentrates on speaker-specific behaviour in larger, suprasegmental units, such as intonation contours. The important role of fundamental frequency for speaker identity is shown. The long vowel /a/ appears to be a very speaker-specific phoneme of Dutch. The importance of cross-validation with multiple recording sessions over time is highlighted in order to achieve more reliable results.

1. Inleiding

Stemmen verschillen; dat spreekt vanzelf. We maken het allemaal wel eens mee dat we aan de telefoon iemands stem herkennen voordat er een naam heeft geklonken. Het simpele woordje 'hallo' is soms al voldoende. Een gesproken uiting bevat kennelijk niet alleen inhoudelijke informatie, maar ook informatie omtrent de identiteit van de spreker.

Hoewel sprekerherkenning mensen behoorlijk afgaat, is het voor computers vaak een fors probleem. In de spraaktechnologie wordt op het moment hard aan dit probleem gewerkt, omdat hier sprake is van een enorme markt (zie Boves & Koolwaaij, te verschijnen).

Politie en justitie zouden graag over machinale hulpmiddelen beschikken waarmee het mogelijk is om een spraakopname van bijvoorbeeld een afperser te analyseren en op grond daarvan de identiteit van de dader vast te stellen (Hollien, 1990). Banken en andere instanties zijn vooral geïnteresseerd in mogelijkheden om de stem als authorisatiemiddel voor bepaalde financiële transacties in te zetten (Furui, 1994).

Het is belangrijk hier te benadrukken dat de stem bepaald niet zo uniek is als een vingerafdruk. Een vingerafdruk is een statisch en onveranderlijk gegeven, terwijl spraak voortdurend verandert in de tijd en sterk beïnvloed wordt door een groot aantal factoren (zie ook sectie 2). Angstvisioenen zoals opgeroepen in Solzjenitsyns roman "In de eerste cirkel", waarin de stem als een unieke vocale vingerafdruk wordt voorgesteld, met alle KGB-achtige gevolgen vandien, stroken beslist niet met de huidige wetenschappelijke inzichten en bevindingen. Dit neemt niet weg dat in spraaktechnologische toepassingen met een beperkte groep coöperatieve sprekers opvallend hoge herkenningsscores kunnen worden bereikt.

Binnen de spraaktechnologie wordt het spraaksignaal, veelal gefilterd door het telefoonkanaal, in zijn geheel gebruikt als invoer om sprekerherkenningsscores te bepalen. De vraag waar precies de sprekerinformatie in de spraak is opgeslagen is daarbij van ondergeschikt belang. Die kwestie behoort eerder tot het werkterrein van de fonetiek.

Wanneer men een blik werpt op datgene waarmee de fonetiek zich in de afgelopen decennia heeft beziggehouden, dan moet men echter vaststellen dat het onderzoek naar sprekerkenmerken nauwelijks aan bod is gekomen. Hoewel een fonoloog als Trubetzkoy dit soort onderzoek expliciet als een werkveld voor de fonetiek bestempelde (Trubetzkoy, 1939: 12), is onder invloed van de Amerikaanse

structuralistische en generatieve traditie de studie van sprekerkenmerken in de fonetiek op een laag pitje komen te staan. Het onderzoek richtte zich in hoofdzaak op de invariante aspecten van de eigenschappen van het spraaksignaal, d.w.z. die eigenschappen die in een linguïstische eenheid hetzelfde blijven, ongeacht de spreker. Kenmerkend is in dit verband de uitspraak van Chomsky en Halle over fonetiek: "phonetics is concerned with grammatically determined aspects of the speech signal" (Chomsky en Halle, 1968: 294).

De laatste jaren is de belangstelling voor individuele aspecten van spraak steeds meer van de grond gekomen (Brown, 1982; Nolan, 1983; Hollien, 1990). Een verklaring hiervoor is ongetwijfeld gelegen in de problemen die zich voordoen bij pogingen om echt invariante kenmerken van fonemen in het spraaksignaal op het spoor te komen. De variatie in foneemrealisaties die wordt geïntroduceerd door de sprekers (en nog een groot aantal andere factoren) is omvangrijker en weerbarstiger dan aanvankelijk werd gedacht.

In dit artikel doen wij verslag van twee fonetisch georiënteerde promotie-onderzoeken naar sprekerkenmerken die wij onlangs aan de universiteit van Nijmegen hebben afgerond. In deze onderzoeken werd de sprekerspecifiekheid van een aantal segmentele (Van den Heuvel, 1996) en suprasegmentele (Kraayeveld, 1997) spraakeigenschappen bestudeerd. Segmentele eigenschappen beperken zich tot een enkel foneem. Een foneem is een spraakklank. Zo bestaat het woord 'goed' uit drie fonemen: /x/, /u/ en /t/. Suprasegmentele kenmerken hebben dan betrekking op grotere eenheden, te beginnen met de syllabe, en betreffen met name variaties in toonhoogte, luidheid, tempo en ritme (Crystal, 1985; Nolan, 1983: 32).

De fonetische aard van de beschreven studies impliceert dat het niet alleen de bedoeling was om optimale sprekerherkenningsresultaten te behalen (dat ligt veel eerder in het belangstellingsveld van de spraaktechnologie) maar om door onderlinge vergelijkingen van spraakkenmerken vast te stellen in welk kenmerk de meeste sprekergerelateerde informatie te vinden is en welke fonetische verklaring daarvoor gegeven kan worden. Zo hoog mogelijke herkenningsscores zijn door ons dus niet nagestreefd. De nadruk lag in eerste instantie steeds op het verwerven van inzicht in de sprekerspecifiekheid van verschillende aspecten van het spraaksignaal.

Dit artikel kan in zekere zin worden beschouwd als een vervolg op het artikel van Rietveld dat eerder in Gramma verscheen (Rietveld, 1988). Een derde artikel, waarin zal worden ingegaan op de spraaktechnologische toepassingsmogelijkheden van sprekerkenmerken, is in voorbereiding (Boves & Koolwaaij, in voorbereiding).

Dit artikel is als volgt opgebouwd. Allereerst gaan we nader in op de vraag wat nu eigenlijk sprekerkenmerken zijn en op welk niveau ze in de spraak gelokaliseerd kunnen worden (sectie 2). Vervolgens vatten we het onderzoek naar suprasegmentele sprekerkenmerken samen dat door Kraayeveld (1997) is uitgevoerd (sectie 3) en daarna bespreken we het door Van den Heuvel (1996) uitgevoerde onderzoek naar segmentele sprekerkenmerken (sectie 4). Tenslotte volgen nog enkele concluderende opmerkingen (sectie 5).

2. Wat zijn sprekerkenmerken?

Bij sprekerkenmerken denkt men in eerste instantie aan interpretaties die een luisteraar geeft aan een stem. Zo kan een stem bijvoorbeeld vriendelijk, kil, of nerveus klinken. Het probleem met dit soort eigenschappen is, dat zij moeilijk te gebruiken zijn, omdat zij per luisteraar kunnen verschillen en moeilijk uit te drukken zijn in maten waarop sprekers direct met elkaar vergeleken kunnen worden. Het is

eenvoudiger onderzoek te doen met behulp van objectief meetbare maten, ook al omdat voor een veranderlijk fenomeen als de stem heel wat spraakmateriaal bestudeerd moet worden, en het verrichten van objectieve metingen in het spraaksignaal vaak veel praktischer, en dus sneller, gaat dan het verzamelen van luisteraarsoordelen. In het onderzoek dat in deze bijdrage beschreven wordt is dan ook geen gebruik gemaakt van luisteraarsoordelen.

Sprekerkenmerken kunnen op allerlei niveaus in de taal onderscheiden worden. Een aantal van die niveaus zijn:

- Gespreksonderwerp; sommige mensen berijden altijd dezelfde stokpaardjes. Zo praten zij over hun kinderen, gezondheidsproblemen, enzovoort.
- Zinsstructuur; sommige mensen onderbreken zichzelf heel vaak, anderen gebruiken vaste formules. Zo verdeelde de vroegere premier Den Uyl problemen altijd in 'twee dingen', met een 'ene kant' en een 'andere kant'.
- Woordkeuze; denk hier aan het gebruik van woorden als 'te gek', of het gebruik van schrijftaalwoorden als 'overigens'.
- Morfeemkeuze; denk hierbij aan het veelvuldig gebruik van verkleinwoordjes of het modevoorvoegsel 'kei-' dat door veel jongeren in het zuiden van Nederland wordt gebruikt.

Dit soort sprekerkenmerken vormt een boeiend onderwerp en de bestudering van dergelijke sprekergebonden stijlverschillen verdient zeker aandacht. Het hier beschreven onderzoek gaat echter over spraakeigenschappen op een wat basaler niveau, namelijk akoestisch-fonetische kenmerken.

Een ander punt dat moet worden belicht is het effect van allerlei 'secundaire' kenmerken op de spraak van individuele sprekers. Sprekerkenmerken zijn deels het gevolg van werkelijk sprekerspecifieke eigenschappen, zoals de unieke anatomische en fysiologische structuur van de spraakorganen, en deels van gedragspatronen die gerelateerd zijn aan groepsgebonden eigenschappen als geslacht, leeftijd, sociaal-economische achtergrond, enzovoort. De invloed van dergelijke sprekereigenschappen, bijvoorbeeld geslacht en leeftijd, kan zelf ook weer veroorzaakt worden door zowel anatomisch-fysiologische verschillen als door sociaal-culturele factoren. Het is moeilijk een grens te trekken om aan te geven welke eigenschappen gezien moeten worden als sprekerspecifiek, en welke door secundaire kenmerken beïnvloed worden. De invloed van sprekerkenmerken die niet gezien worden als 'echt' sprekerspecifiek (bijvoorbeeld geslacht en leeftijd) op de te onderzoeken variabelen kan worden gecontroleerd door deze factoren systematisch te variëren. Voor de factor geslacht moeten dan zowel mannen als vrouwen worden onderzocht, waarna mannen ook alleen binnen de mannengroep en vrouwen alleen binnen de vrouwengroep bestudeerd worden.

Een veelgekozen benadering in het onderzoek naar sprekerkenmerken is het onderverdelen van de spraakkenmerken in 'organische' en 'aangeleerde' kenmerken (zie bijvoorbeeld Wolf, 1972). Daarbij wordt de, soms impliciete, aanname gedaan dat organische kenmerken minder aan variatie onderhevig zijn, en dus beter te gebruiken zijn om sprekers te onderscheiden. Deze benadering valt uiteindelijk echter niet vol te houden, omdat het in feite niet mogelijk is organische en aangeleerde kenmerken te onderscheiden (Nolan, 1983). Een kenmerk als gemiddelde toonhoogte, bijvoorbeeld, wordt voor een belangrijk deel bepaald door de bouw van de stembanden. Binnen de bandbreedte aan mogelijke toonhoogten die door de bouw van de stembanden bepaald wordt, is echter een grote verscheidenheid aan

toonhoogten mogelijk. De daadwerkelijk gebruikte toonhoogte wordt door een veelheid aan sociale en psychische factoren bepaald.

Ter verkrijging van een maat voor sprekerspecifiekheid van spraakkenmerken moeten twee bronnen van variabiliteit in beschouwing worden genomen. Aan de ene kant bestaat er uiteraard variatie tussen sprekers. Er bestaat echter ook variatie binnen de uitingen van één en dezelfde spreker. Pas wanneer voor een parameter de verschillen tussen de sprekers beduidend groter zijn dan de variatie die kan worden aangetroffen in verschillende uitingen van één spreker, kan worden gesteld dat deze parameter sprekerspecifiek is. Daarom wordt het quotiënt van tussensprekervariatie en binnensprekervariatie in sprekeronderzoek vaak als maat voor sprekerspecifiekheid aangehouden (zie bijvoorbeeld Wolf, 1972; Bonastre, Meloni en Langlais, 1991).

3. Sprekerkenmerken op suprasegmenteel niveau

In Kraayeveld (1997) wordt een onderzoek gepresenteerd dat tot doel had vast te stellen in hoeverre suprasegmentele parameters kunnen bijdragen aan correcte sprekerherkenning. Suprasementele kenmerken hebben betrekking op spraakeenheden die langer duren dan een foneem (zie sectie 1).

In het onderzoek werden twee typen suprasegmentele parameters gedefinieerd. Het eerste type betrof maten die verkregen werden door over een bepaald tijdsinterval te middelen. Deze werden gemiddelde maten genoemd. Een voorbeeld van een dergelijke maat is de spreeksnelheid, waarbij over een bepaald tijdsinterval het aantal uitgesproken lettergrepen geteld werd, hetgeen vervolgens gedeeld werd door de duur van het betreffende interval.

Het tweede type suprasegmentele parameters werd gemeten op bepaalde punten in een uiting: aan het begin en eind van de uitingen, en op *keerpunten* in de toonhoogtecontour, dat wil zeggen op punten waar toonhoogtebewegingen beginnen of eindigen. Omdat de realisaties van toonhoogtebewegingen door verschillende sprekers alleen vergeleken kunnen worden op vaste punten in vaste uitingen/toonhoogtecontouren werden deze maten contourgebonden maten genoemd.

Voor de twee parametertypen, de gemiddelde en de contourgebonden maten, wilden we vaststellen in hoeverre zij apart of in combinatie gebruikt kunnen worden om sprekers te identificeren.

3.1 Methode

Voor de meting van de gemiddelde maten vergaarde Kraayeveld spraakfragmenten van 15 seconden uit twee taken: een interview en een voorleestaak. Er werden 1000 fragmenten gebruikt: vijf spontane fragmenten (uit het interview) en vijf voorgelezen fragmenten (uit de voorleestaak) van 50 sprekers, verzameld in twee opnamesessies, waartussen een tijdsinterval van zeven maanden lag.

De gemiddelde maten kunnen worden verdeeld in drie groepen: toonhoogtematen, amplitudematen en temporele maten. De gebruikte toonhoogtematen waren de gemiddelde toonhoogte, de variabiliteit van de toonhoogte, en twee maten voor de duurvariatie van dicht bij elkaar gelegen perioden, oftewel toonhoogte*perturbatiematen*. Ook voor de amplitude werden de variabiliteit en twee perturbatiematen bepaald. Er werden drie temporele maten gebruikt: de articulatiesnelheid, de

hoeveelheid pauze en het percentage stemhebbende spraak. In totaal waren er aldus 10 verschillende gemiddelde maten.

Op het gebied van de contourgebonden maten werden, zoals hierboven aangegeven, metingen verricht op verschillende posities in de toonhoogtecontour. Deze metingen bestonden in de eerste plaats uit toonhoogtemetingen. De eindtoonhoogte van de uiting werd uitgedrukt in hertz, de andere metingen in het aantal semitonen verschil met deze eindtoonhoogte (een semitoon is gelijk aan de toonafstand tussen een witte en een zwarte toets op de piano). Ook de duur van de toonhoogtebewegingen werd gemeten. Tenslotte werd voor alle toonhoogtebewegingen het tijdsinterval tussen het begin van de beweging en het begin van de klinker in de betrokken lettergreep vastgesteld (*synchronisatietijd*). Met deze maat werd een brug geslagen tussen de segmentele en de suprasegmentele structuur van de uitingen.

Voor het meten van contourgebonden parameters hadden wij zinnen nodig die bij iedere realisatie een identiek toonhoogtecontour zouden opleveren. We vonden zinnen met dergelijke eigenschappen in de wereld van de sport; we gebruikten opnamen van drie zinnen van de vorm *De leren wonnen van de Denen met drieëén*. Omdat de zinnen in twee opnamesessies met 50 sprekers verzameld werden bestond het gegevensbestand uiteindelijk uit 300 zinnen. Metingen werden verricht aan de stijging op de eerste lettergreep van de eerste nationaliteit, aan de stijging op het eerste getal van de score en aan de daling op het tweede getal van de score. Het totale aantal contourgebonden maten dat we per zin gebruikten was 20.

De gecombineerde sprekeridentificatie-mogelijkheden van zowel de gemiddelde als de contourgebonden maten werden vastgesteld met behulp van discriminantanalyses. Door middel van deze techniek werd gepoogd sprekers optimaal van elkaar te onderscheiden op basis van de invoervariabelen, en deze sprekeronderscheiding te evalueren in termen van sprekerherkenningspercentages. Hierbij speelt de verhouding van tussen- en binnensprekervariatie een essentiële rol. Voor nadere uitleg over discriminantanalyses verwijzen wij naar Klecka (1980) en Stevens (1986).

3.2 Resultaten en conclusies

In het eerste deel van het onderzoek werden gemiddelde maten bepaald in wat langere spraakfragmenten (15 seconden). In een discriminantanalyse met alle fragmenten vonden we een genormaliseerde identificatiescore van 60%, hetgeen betekent dat boven de hoeveelheid sprekerherkenning die al door toeval zou worden verkregen (bij 50 sprekers is dat 2%) nog eens 60% herkend werd. Het percentage correcte herkenning bleek te kunnen worden verhoogd door gegevens van alleen mannen (65% correcte sprekeridentificatie) of van alleen vrouwen (63% correcte identificatie) te analyseren. Een nog grotere toename in de sprekeridentificatiescore kon worden verkregen door alleen voorgelezen (88% correcte identificatie) of alleen spontane spraakfragmenten (70% correcte identificatie) te analyseren. Vooral voorgelezen fragmenten bleken aldus zeer sprekerspecifiek.

In alle genoemde analyses werd een vooraanstaande rol gespeeld door de gemiddelde toonhoogte, maar sprekerherkenning bleek niet onmogelijk te worden zonder deze maat (48% correcte identificatie), terwijl een discriminantanalyse met alleen de gemiddelde toonhoogte als predictormaat weinig succesvol was (8% correcte identificatie).

Sprekerspecifieke maten hebben weinig praktische betekenis als sprekers na verloop van tijd heel andere waarden realiseren. Daarom werd in dit onderzoek veel belang gehecht aan de zogenaamde kruisvalidatie-analyses. Daarin werd spraakmateriaal uit één van de opnamesessies gebruikt om discriminantfuncties te bepalen die vervolgens werden toegepast om het spraakmateriaal uit de andere sessie aan de sprekers toe te wijzen. Helaas bleek de identificatiescore bij kruisvalidatie met alle 10 gemiddelde maten slechts 33% te bedragen. In uitsluitend voorgelezen materiaal lag dit percentage hoger: 54%.

Het doel van het tweede deel van het onderzoek was voornamelijk het vaststellen van de sprekeridentificerende eigenschappen van de contourgebonden maten. Bovendien werd het identificerende vermogen van de contourgebonden maten vergeleken met dat van de gemiddelde maten en van de combinatie van de twee typen.

In dit tweede deel van het onderzoek gingen we op dezelfde manier te werk als in het eerste. Nu werden enkel de 300 sportzinnen geanalyseerd. Eerst werd het sprekeronderscheidend vermogen van de gemiddelde en de contourgebonden maten vastgesteld in discriminantanalyses die werden uitgevoerd op zowel het gehele materiaal als op deelverzamelingen ervan. Met gebruikmaking van alle 20 contourgebonden maten werd in het totale zinnenmateriaal een score van 86% correcte sprekerherkenning bereikt.

Bij het vergelijken van de analyses met contourgebonden maten met de analyses met gemiddelde maten (of met de combinatie van beide soorten maten) moet rekening gehouden worden met het feit dat met een groter aantal parameters betere resultaten kunnen worden geboekt. Daarom werden voor de vergelijking van de verschillende soorten maten analyses verricht waarbij alleen de 10 meest succesvolle parameters tot de analyses werden toegelaten. De herkenningsscore voor de contourgebonden maten lag in deze analyses iets lager dan die voor de gemiddelde maten: 77% voor de contourgebonden maten en 81% voor de gemiddelde maten. Gecombineerd was de score niet veel hoger: 84%.

In zowel de discriminantanalyse met de gemiddelde maten als in de analyse met beide typen parameters was de gemiddelde toonhoogte nog steeds de belangrijkste sprekerherkenningsvariabele. In de analyses met de contourgebonden maten werd een vergelijkbaar belangrijke rol gespeeld door de eindtoonhoogte van de zin. Net als voor de gemiddelde toonhoogte vonden we dat het een belangrijke maat was, maar dat sprekeridentificatie er niet volledig van afhankelijk is.

Bij de contourgebonden maten bleek uitsplitsen naar geslacht geen verbetering op te leveren: de score voor mannen was daar 76%, voor vrouwen 86% en in de gecombineerde analyse 77%. Ook bij de contourgebonden maten waren de identificatiescores in de kruisvalidatie-analyses laag: 42% correct voor zowel de contourgebonden als de gemiddelde maten en 44% voor de combinatie van deze twee typen maten. Ook met de in de sportzinnen verkregen gegevens lijkt toepassing van (alleen) suprasegmentele maten in praktische situaties dus niet bijzonder kansrijk.

De eindconclusie die op basis van de gerapporteerde gegevens getrokken kan worden is dat de gebruikte suprasegmentele maten wellicht een bruikbare bijdrage kunnen leveren aan de herkenning van sprekers, maar vanwege de lage identificatiescores in de zo belangrijke kruisvalidaties lijken de gebruikte supra-

segmentele maten alleen in aanmerking te komen voor een ondersteunende, aanvullende rol.

4. Sprekerkenmerken op segmenteel niveau

In Van den Heuvel (1996) wordt een aantal experimenten gepresenteerd dat ten doel had verschillen in sprekerspecifiekheid van Nederlandse foneemrealisaties op te sporen en een fonetische verklaring voor die verschillen te vinden. Sprekerspecifiekheid is hierbij opgevat als het quotiënt van tussensprekervariatie en binnensprekervariatie, zoals eerder werd aangegeven in sectie 2.

Er werd gekeken naar sprekergerelateerde verschillen in *duur* en in *spectrale samenstelling* van foneemrealisaties. De duur van een foneem wordt uitgedrukt in milliseconden. De spectrale samenstelling van een foneem wordt bepaald door te berekenen welke clusters van frequenties (boventonen) het sterkst in het spraaksignaal vertegenwoordigd zijn gedurende een tijdsinterval (frame). Zo'n cluster van versterkte frequentiecomponenten wordt een formant genoemd, een term die hieronder regelmatig zal terugkeren.

De duur en de spectrale samenstelling van een foneemrealisatie wordt in belangrijke mate bepaald door naburige fonemen. Deze invloed noemen we coarticulatie. Wanneer iemand bijvoorbeeld het woordje 'stroop' uitspreekt, dan zullen de lippen bij de realisatie van /s/ al getuit zijn in afwachting van de nog komende geronde /o/. Bij wijze van controle kan de lezer bij zichzelf vast stellen dat het tuiten van de /s/ achterwege blijft, indien in plaats van 'stroop' het woord 'streep' wordt gezegd. In het onderzoek is nagegaan in hoeverre coarticulatie (de invloed van het ene foneem op de realisatie van een naburig foneem) sprekerspecifiek is.

Binnen het kader van dit artikel kan uiteraard niet meer dan een korte samenvatting van het onderzoek worden geboden. De lezer wordt verwezen naar Van den Heuvel (1996) voor een gedetailleerd verslag.

4.1 Methode

Voor het onderzoek zijn twee experimenten uitgevoerd. In het eerste experiment werden de foneemduren onderzocht; aan dit experiment deden vijf mannen en vijf vrouwen mee. In het tweede experiment werden de spectra van de fonemen onderzocht; aan dat experiment deden 15 mannen mee. Door een panel van vijf logopedisten werden deze sprekers uit een grotere groep geselecteerd op grond van het criterium dat ze standaard Nederlands spraken en geen pathologische bijzonderheden vertoonden.

In beide experimenten bestond het spraakmateriaal uit 24 pseudowoorden. Elk woord was opgebouwd uit een beginconsonant C_1, een vocaal V, een tweede consonant C_2 en een schwa ('stomme e'). In elk woord was V één van de klinkers /a/, /i/ of /u/ en C_1 en C_2 waren steeds geselecteerd uit de set /p, t, k, d, s, m, n, r/. Voorbeelden van zulke woorden zijn: 'poede', 'nare' en 'kiene'.

Elk woord werd 10 maal door elke spreker uitgesproken, zonder inbedding in een draagzin. Voor het eerste experiment resulteerde dat in 2400 woordrealisaties (24 woorden x 10 sprekers x 10 herhalingen) en voor het tweede experiment derhalve in 3600 woordrealisaties.

In experiment 1 werd voor elk foneem de duur bepaald. Voor experiment 2 werd van elk van de fonemen /a, i, u, m, n, s/ een frame uit het midden als spectrale

representant van dat foneem gekozen.

Via een score-model benadering werden vergelijkingen opgesteld voor tussen- (INTER) en binnensprekervariatie (INTRA) in duren en spectra van de fonemen. Vervolgens werd een sprekerspecifiekheidsindex SSI gedefinieerd: INTER gedeeld door INTRA (vgl. sectie 2 hierboven). Ook werd er via dit scoremodel een coarticulatiemaat (COART) opgesteld. De COART-maat kwantificeert de afstand tussen een in isolatie gerealiseerd foneem en een foneem dat in een specifieke foneemomgeving wordt gerealiseerd. Hoe groter deze afstand, hoe groter de coarticulatie.

In dit onderzoek is eveneens gebruik gemaakt van discriminantanalyses (zie ook sectie 3). Door middel van deze techniek werd gepoogd sprekers optimaal van elkaar te onderscheiden op basis van hun foneemspectra, en deze spreker-onderscheiding te evalueren in termen van sprekerherkenningspercentages. Ook hierbij speelt de verhouding van tussen- en binnensprekervariatie een essentiële rol.

4.2 Resultaten foneemduren

Allereerst werden de foneemduren van experiment 1 geanalyseerd. Uit een variantie-analyse bleek dat de factoren vocaalidentiteit en consonantomgeving een aanmerkelijk sterker effect op de vocaalduren hadden dan factoren die met de sprekers te maken hadden. Verder bleek uit de waarden van SSI dat de duren van de (lange) vocalen het meest sprekerspecifiek waren.

De gevonden sprekerspecifiekheid van de vocalen kon voor een zeer belangrijk deel aan individuele verschillen in *spreektempo* worden toegeschreven. Dit wordt in figuur 1 geïllustreerd. Hier worden per foneem twee SSI-waarden getoond; in de witte balken is de invloed van het individuele spreektempo behouden en in de gekleurde balken is deze invloed verwijderd. De SSI-waarden van /p/, /t/ en /k/ aan het begin van een woord konden niet bepaald worden; de SSI-waarde van /r/ in C_2-positie is onbetrouwbaar.

Fig. 1 ongeveer hier:
Gemiddelde SSI-waarden voor foneemklassen (C_1, V, C_2) en afzonderlijke fonemen, met spreektemponormalisatie (gearceerde staven) en zonder spreektemponormalisatie (witte staven). De consonanten zijn opgesplitst naar C_1- (staaf links) en C_2 positie (staaf rechts).

Vervolgens werd de invloed van de consonantomgeving op de vocaalduur onderzocht. Deze invloed werd opgevat als een vorm van coarticulatie. De grootste vocaalverlenging werd gevonden ten gevolge van de postvocale /r/ (zoals in 'nare'); een wat minder grote verlenging werd vastgesteld ten gevolge van de prevocale /r/ (zoals in 'rade').

De sprekerafhankelijkheid in de coarticulatie van /a/ door /r/ werd aan een nader onderzoek onderworpen. Het bleek dat de verlenging van /a/ door *post*vocale /r/ kon worden geïnterpreteerd als een verplichte, sprekeronafhankelijke, regel en de verlenging van /a/ door de *pre*vocale /r/ als een optionele, sprekerafhankelijke, regel (zie ook Van den Heuvel, Rietveld en Cranen, 1994).

4.3 Resultaten spectra

Het spraakmateriaal uit het tweede experiment werd gebruikt om de sprekerspecifiekheid van de foneem*spectra* te onderzoeken. Dit onderzoek werd uitgevoerd met behulp van discriminantanalyses.

De spectra van de stationaire fonemen in het spraakmateriaal (/a, i, u, m, n,

s/) werden geanalyseerd. In termen van sprekerherkenningspercentages werd de volgende rangorde voor de fonemen gevonden:

/a/ > /n/ > /i/ > /m/ > /u/ > /s/.

Met andere woorden: de /a/ was het meest sprekerspecifiek en de /s/ het minst. De gevonden volgorde levert een eenvoudig te onthouden woord op.

Bij nadere beschouwing bleek de tweede formant de grootste bijdrage aan de sprekerherkenning van /a/ en /i/ te leveren, terwijl deze formant het minst bijdroeg aan de herkenning van /u/ en /m/. Voor /u/ en /m/ is dit resultaat toe te schrijven aan coarticulatieverschijnselen. Omdat het spectrum van de /u/ en de /m/ (vooral hun tweede formant) afhankelijk is van de omliggende fonemen ontstaat er een grote spreiding in de spectra van een spreker. Als gevolg daarvan vertonen de spectra van verschillende sprekers een grotere neiging om elkaar te overlappen en hierdoor kunnen de sprekers moeilijker uit elkaar worden gehouden.

De filterbanden met de grootste sprekerspecifiekheid waren vaak gelokaliseerd rond de formantfrequenties. Het daaruit volgende vermoeden dat de bandbreedten van de formanten wellicht meer sprekerspecifiek waren dan de formantfrequenties zelf, werd bij nadere analyse niet bevestigd.

Tenslotte werd de sprekerspecifiekheid van de coarticulatie in de eerste drie formanten van /a/, /i/ en /u/ onderzocht. Hiervoor werd de COART-maat gebruikt. De sterkste coarticulatie werd gevonden in /u/. Vooral de tweede formant van /u/ bleek sterk door de nasale (/m/ en /n/) en de alveolaire (/n/, /d/ en /t/) consonanten in C_1-positie beïnvloed te worden.

De coarticulatie van /a/ en vooral /u/ in de verschillende consonantomgevingen bleek significant sprekerafhankelijk te zijn. Het is daarom mogelijk dat de coarticulatie in de vocalen kan helpen sprekers te identificeren. Dit werd uitgetest door eerst te proberen sprekers te herkennen op basis van de eerste drie vocaalformanten en vervolgens de coarticulatie-informatie (COART) hieraan toe te voegen en de herkenning opnieuw op te starten. Ook hier werd discriminantanalyse gebruikt om de sprekers zo goed mogelijk te groeperen en te herkennen. Het bleek dat de sprekerherkenningsscores door toevoeging van de coarticulatie-informatie niet of nauwelijks verbeterden. Hoewel de sprekers op basis van COART beter herkend werden dan op grond van het toeval viel te verwachten waren de formanten als zodanig betere voorspellers van sprekeridentiteit.

5. Besluit

Sprekerinformatie en linguïstische informatie worden niet in verschillende delen van de spraak aangetroffen. Vaak komen zij voor in dezelfde delen van het spectrum of van de intonatiecontour. Op het segmentele niveau bleek dit uit de observatie dat de meest sprekerspecifieke frequenties in de foneemspectra zich op en rond de formantfrequenties bevinden; dit zijn precies de frequenties die de identiteit van een vocaal bepalen. Op suprasegmenteel niveau bleek de toonhoogte aan het einde van de uiting zeer sprekerspecifiek, terwijl deze locatie tevens van groot belang is voor de bepaling van het zinstype (vraagzin of niet).

Een deel van de onderzochte verschijnselen bleek sprekerafhankelijk te zijn, bijvoorbeeld de verlenging van /a/ door prevocale /r/, de coarticulatie van vocalen in

het spectrale domein en de synchronisatie van de segmentele en de suprasegmentele laag. Dit geeft aan dat fonetische verschijnselen niet slechts bij één of bij enkele sprekers onderzocht moeten worden, maar bij een veel groter aantal, omdat men anders het gevaar loopt sprekergebonden effecten aan te zien voor algemeen fonetisch-linguïstische effecten.

Onze data toonden een positief verband tussen de sprekerspecifiekheid van een foneem, zowel in duur als in spectrale samenstelling, en zijn frequentie van voorkomen. In het Nederlands komt namelijk de /a/ het meest voor en de /u/ het minst. Zwakkere, maar soortgelijke verbanden werden voor de consonanten gevonden. Mogelijk worden de realisaties van vaak voorkomende spraakgebeurtenissen wat slordiger, waardoor sprekerkarakteristieken gemakkelijker aan de dag kunnen treden.

In het onderzoek naar sprekerspecificiteit in suprasegmentele spraakkenmerken werd getest in hoeverre de gegevens van de ene opnamesessie voorspellende waarde hebben voor gegevens uit een volgende sessie. Daardoor gaf het onderzoek meer duidelijkheid over de bruikbaarheid van die kenmerken voor sprekerherkenning dan het onderzoek naar segmentele eigenschappen; de suprasegmentele eigenschappen lijken een interessante aanvullende rol te kunnen vervullen bij sprekerherkenning, terwijl sprekerherkenning op basis van alleen deze kenmerken niet erg kansrijk lijkt.

Op zich werden ook voor de segmentele eigenschappen redelijke sprekerherkenningsscores gevonden, maar hierbij ontbrak de hertest met materiaal dat op een ander tijdstip was vergaard. Het lijkt interessant nader te onderzoeken in hoeverre een combinatie van segmentele en suprasegmentele eigenschappen tot succesvolle toepassing kan leiden.

De toepassingsmogelijkheden van spraakdata kunnen mogelijk verder worden vergroot door referentiemateriaal te verzamelen in meerdere opnamesessies. Dan zal de variatie tussen opnamesessies al in het referentiemateriaal verdisconteerd worden, waardoor meer realistische sprekerherkenningspercentages kunnen worden bereikt.

Henk van den Heuvel
Vakgroep Taal & Spraak
Katholieke Universiteit Nijmegen
Postbus 9103
6500 HD Nijmegen
tel (024) 3616087
fax (024) 3615939
H.v.d.Heuvel@let.kun.nl

Hans Kraayeveld
Delta Lloyd Verzekeringsgroep N.V.
Informatica Service, team Netwerkbeheer
Postbus 1000
1000 BA Amsterdam
tel (020) 5943604
fax (020) 5942080
Kraayeveld@dl.e-mail.com

Bonastre, J.F., H. Meloni en P. Langlais (1991). Analytical strategy for speaker identification. *Proceedings of EUROSPEECH '91*, Genua: ESCA/IIC, 427-430.

Boves, L. en J. Koolwaaij (in voorbereiding voor GRAMMA/TT)

Brown, R.S. (1982). What is speaker recognition? *Journal of the International Phonetic Association*, 12, 13-24.

Chomsky, N. en M. Halle (1968). *The sound pattern of English*. New York, Evanston, Londen: Harper & Row.

Crystal, D. (1985). *A dictionary of linguistics and phonetics, 2nd ed*. Oxford: Basil Blackwell.

Furui, S. (1994). An overview of speaker recognition technology. *Proceedings of the ESCA workshop on automatic speaker recognition, identification and verification. Martigny*, 1-9.

Heuvel, H. van den (1996). Speaker variability in acoustic properties of Dutch phoneme realisations. Proefschrift, Katholieke Universiteit Nijmegen.

Heuvel, H. van den, T. Rietveld en B. Cranen (1994). Methodological aspects of segment and speaker-related variability. A study of segmental durations in Dutch. *Journal of Phonetics*, 22, 386-406.

Hollien, H. (1990). *The acoustics of crime. The new science of forensic phonetics*. New York, Londen: Plenum Press.

Klecka, W.R. (1980). *Discriminant analysis*. Beverly Hills, Londen: Sage Publications.

Kraayeveld, J. (1997). Idiosyncrasy in prosody. Proefschrift, Katholieke Universiteit Nijmegen.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge, New York, Melbourne: Cambridge University Press.

Rietveld, A.C.M. (1988). *Sprekerherkenning: methoden, mogelijkheden, moeilijkheden*. Gramma, 12, 2, 117-132.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale (NJ): Lawrence Erlbaum.

Trubetzkoy, N.S. (1939). *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague, 7.

Wolf, J.J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51, 2044-2056.