Spraaktechnologie in opmars

Gerrit Bloothooft (versie 15 juni 1998)

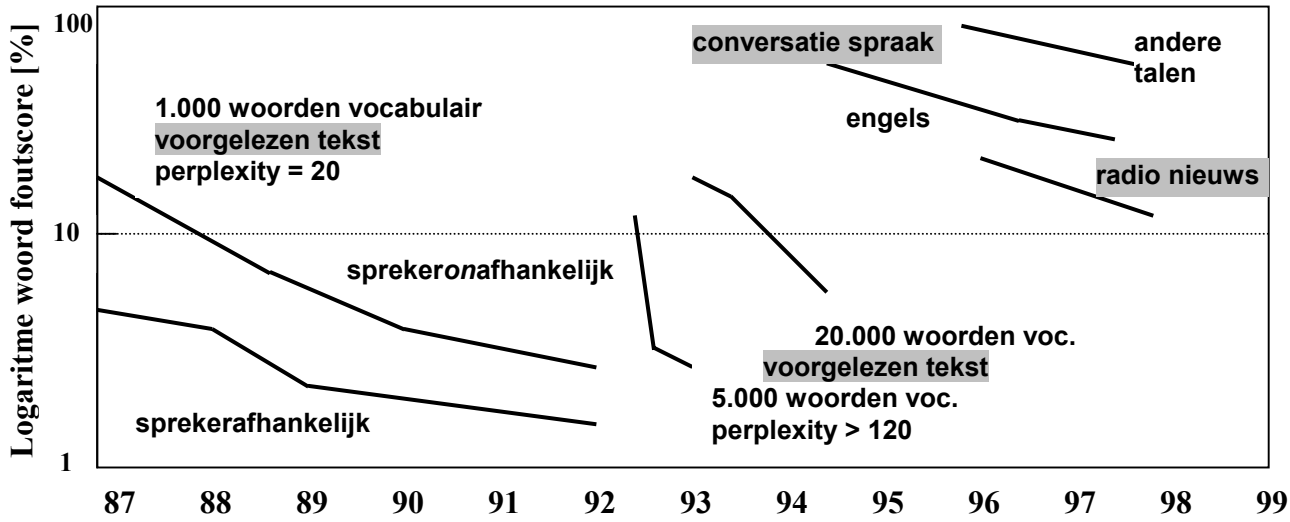
'U spreekt met het reizigersinformatiesysteem van de Nederlandse Spoorwegen. Waarmee kan ik u van dienst zijn?'.
'Ik wil overmorgen met de trein van Utrecht naar Maastricht'.
'Hoe laat wilt u woensdag 3 juni vanuit Utrecht naar Maastricht vertrekken?'.
'Ongeveer om negen uur'.
'Uw trein vertrekt om negen uur vier uit Utrecht, aankomst om elf uur drie in Maastricht. Is er verder nog iets van uw dienst?'.
'Nee, dank u'.

Zo'n korte dialoog kan iedereen meemaken die telefonisch treintijdinformatie (0900-9292) opvraagt. Als het druk is en de wachttijd voor een afhandeling door een informatrice te lang wordt, wordt het gesprek overgenomen door de computer als de beller dat wil. 8000 mensen worden op deze manier dagelijks geconfronteerd met een geavanceerde toepassing van spraak- en taaltechnologie. En dat is nog maar het begin. Wat tien jaar geleden nog toekomstmuziek leek, is nu beschikbaar voor een breed publiek en waarschijnlijk schiet ons voorstellingsvermogen te kort om te bedenken hoe revolutionair spraak- en taaltechnologie - als een fundamenteel onderdeel binnen informatie- en communicatietechnologie - de samenleving in de volgende eeuw zal beïnvloeden.

Veel barrières moesten worden geslecht om het bovenstaande gesprekje mogelijk te maken. Ten eerste moet de computer in staat zijn om lopende spraak te herkennen, vervolgens moet de spraak begrepen en geinterpreteerd worden en tenslotte moet een antwoord of aanvullende vraag met kunstmatige spraak geproduceerd worden. We zullen voor elke fase nagaan wat de huidige stand van zaken is, wat er voor nodig was om die situatie te bereiken en wat ons aan verdere verbeteringen te wachten staat.

Over de interpretatie van de letterreeks 'trein' zal weinig discussie ontstaan. De letters zijn eenduidige symbolen die samen een betekenisvol woord vormen. Bij spraak is er echter sprake van grote variabiliteit in het spraakgeluid waarin de betekenis 'trein' vervat ligt. Elke spreker spreekt 'trein' op eigen manier uit door de eigenschappen van stemplooien en de anatomie van de mond-keelholte. Het woord kan daarnaast snel of langzaam, luid of zacht, met verschillende emoties, of met dialect worden uitgesproken. Een systeem voor automatische spraakherkenning moet (net zoals de mens maar niet noodzakelijk op dezelfde manier) al deze variatie in het spraakgeluid beheersen. Dat lukt alleen als het systeem deze variatie heeft geleerd. Er zijn twee manieren om aan een systeem de benodigde kennis toe te voegen. Ten eerste kan de ontwikkelaar regels maken die de te verwachten variatie in spraak beschrijven. Lange tijd is gedacht dat fonetici deze kennis zouden bezitten omdat enkele experts in staat waren om uit een spectrogram - dat de akoestische eigenschappen van de spraak visueel beschrijft - de gesproken woorden af te leiden. Blijkbaar gebruiken deze fonetici hierbij regels. Het zou voor automatische spraakherkenning alleen maar nodig zijn om deze regels expliciet te maken en in een expertsysteem onder te brengen. De pogingen hiertoe zijn in de tachtiger jaren mislukt. Er is bij mijn weten ook nooit onderzoek gedaan hoe goed de experts feitelijk waren in het lezen van spectrogrammen. Ondanks hun indrukwekkende analytische vaardigheden mag vermoed worden dat er feitelijk sprake zal zijn geweest van een foutenpercentage dat onacceptabel is voor technologische toepassingen.

Een tweede aanpak is om alle mogelijke variatie in spraak in kaart te brengen. Dat kan alleen als er een geschikt model is om deze variatie te beschrijven. Dat model moet in staat zijn om de kans op een bepaalde realisatie van een spraakklank en van de kans van een opeenvolging van spraakklanken tot een woord weer te geven. Het Hidden Markov Model (HMM) is hierbij het dominerende model geworden en HMMs hebben het daarbij in de afgelopen twintig jaar duidelijk gewonnen van neurale netwerken, alhoewel de laatsten soms in hybriede systemen worden ingezet. De kracht van een Hidden Markov Model is ongetwijfeld dat de informatie die erin opgeslagen kan worden gecumuleerd kan worden door ongesuperviseerde training van zeer grote hoeveelheden trainingsmateriaal. Daarnaast is de modellering met HMMs een universele techniek gebleken die op alle beschijvingsniveaus van spraak en taal toepasbaar is. HMMs beschrijven de akoestische variabiliteit van een spraakklank, ze beschrijven hoe spraakklanken aaneenrijgen tot woorden, ze kunnen de prosodie van spraak beschrijven, ze beschrijven hoe woorden aaneenrijgen tot zinnen of tot welke syntactische klasse een woord kan behoren (*part of speech tagging*).



Figuur 1: Percentages woordfouten in verschillende spraakherkennings tests, uitgevoerd door NIST over de laatste 10 jaar (naar Pallett, 1998).

Een goed overzicht van de ontwikkeling in het gebruik van Hidden Markov Modellen in automatische spraakherkenning zijn de resultaten die gedurende de laatste tien jaar zijn verkregen in tests gecoordineerd door het Amerikaanse National Institute of Standards (NIST) (Pallett, 1998). Figuur 1 geeft het percentage foutief herkende woorden in deze tests. De tests zelf laten een interessante progressie zien in de aangevatte taak. Als met een bepaalde taak een acceptabel foutpercentage werd behaald werd voor een nieuwe, veel moeilijker taak gekozen. De tests startten tien jaar geleden op het moment dat met het SPHINX systeem een doorbraak naar de herkenning van lopende spraak was bereikt (Lee et al., 1998). De eerste tests betroffen het herkennen van voorgelezen spraak uit een vocabulaire van 1000 woorden waarbij gewerkt werd met zowel spekerafhankelijke en sprekeronafhankelijke technologie. De voorspelbaarheid van de tekst, uitgedrukt in perplexiteit (ruwweg het aantal mogelijk woordkandidaten dat gemiddeld op een woord volgt), was relatief groot, zo'n 20. In de zes jaar dat deze tests werden uitgevoerd (1988-1992) zakte het percentage foutief herkende woorden van 7% tot 2% (sprekerafhankelijk) en van 20% tot 5% (sprekeronafhankelijk). Vervolgens werden gedurende 1992-1995 tests uitgevoerd met veel grotere vocabulaires (5000 en 20000 woorden) en een kleinere

voorspelbaarheid van de tekst (perplexiteit > 120). Voor de grotere vocabulaires schommelt het percentage woordfouten tegenwoordig rond de 10% voor de verschillende systemen en het lijkt erg moeilijk om daaronder te komen. Het is dan interessant om te weten wat mensen in zo'n taak zouden presteren. Ondanks verschillen tussen mens en machine in de hoeveelheid training en in de afname van de test blijkt dat de mens gemakkelijk onder 3% scoort (van Leeuwen et al., 1995). En ook niet-moedertaal-sprekers van het Engels (maar met een redelijke beheersing ervan) doen het nog beter dan de beste spraakherkenner.

Toen met voorgelezen spraak redelijke resultaten werden behaald, wat van belang is voor de ontwikkeling van bijvoorbeeld de spraakgestuurde tekstverwerkers of voor spraakgestuurde computers, verschoof de aandacht naar de veel moeilijker taak van spontane spraak. Het is interessant dat er in de laatste tien jaar een brede belangstelling is gekomen voor spontane spraak. Vanuit de spraaktechnologie was dat een logische stap die volgde op een ontwikkeling van het automatisch herkennen van los uitgesproken woorden tot voorgelezen continue spraak met steeds grotere vocabulaires. Daarnaast werd ook in het fonetisch onderzoek onderkent dat onze kennis van spraakproductie en spraakperceptie vaak gebaseerd was op opnamen die onder kunstmatige omstandigheden werden gemaakt en dat een verschuiving naar onderzoek aan meer natuurlijker opnamen gewenst en zelfs noodzakelijk was. Overigens heeft deze opmerkelijke parallelle ontwikkeling niet geleid tot onderlinge kruisbestuiving van resultaten. Wel zijn er grote spraakbestanden beschikbaar gekomen waar ook fundamenteel onderzoek aan verricht zou kunnen worden.

In de NIST tests moest de computer spontane telefoonconversaties herkennen. Het zal niet verbazen dat hoge percentages woordfouten tot 70% resulteerden. Spontane spraak is slordig, zowel in articulatorische realisaties als in tekstopbouw, terwijl ook de kwaliteit van de telefoonlijn de herkenning niet gemakkelijk maakt. Slechts geleidelijke verbeteringen tot 30% zijn de laatste jaren bereikt. Naast deze test ligt de laatste jaren de aandacht bij ander interessant materiaal: audiobanden uit archieven van radio en televisie. Om hieruit interessante passages te halen moet het hele materiaal door automatische spraakherkenning in kaart zijn gebracht. De opnamen kan voorgelezen en spontane spraak bevatten van allerlei kwaliteit. Er kan achtergrondmuziek bij zitten, of achtergrondruis, wat meestal nog katastrofaal is voor huidige spraakherkenners. De eerste tests gaven net zoals bij telefoonconversaties zeer hoge foutenpercentages, rond de 70% maar inmiddels is men al erg blij met percentages rond 30%. Dat lijkt hoog, één op de drie woorden wordt foutief herkend. Maar dit soort percentages krijgt pas betekenis in de context van een toepassing. Als het de bedoeling is om uit tienduizenden uren radioarchief de minuten te halen toen er over een bepaald onderwerp werd gesproken dan is perfecte automatische spraakherkenning helemaal niet nodig. Er wordt op trefwoorden gezocht en in de analyse van een bepaald gesprek mag worden aangenomen dat een trefwoord wel eens een keer goed herkend zal worden. Dat is voldoende om de gewenste passages te traceren.

De voortdurende progressie in de resultaten van automatische spraakherkenning vooral te maken met een enorme toename in de hoeveelheid trainingsmateriaal dat wordt gebruikt. Er is in de negentiger jaren een geweldige inspanning gestart om spraak- en taalmateriaal te verzamelen. Dit viel gelukkigerwijs samen met de ontwikkeling van CD-ROMs voor goedkope opslag en brede verspreiding. Het beschikbaar komen van grote spraak- en tekstbestanden zal doorzetten en mogelijk een ongekende invloed hebben op technologisch maar ook fundamenteel fonetisch en taalkundig onderzoek. Verschijnselen kunnen nu immers op veel grotere schaal onderzocht worden. Het Linguistic Data Consortium (LDC) in

Amerika heeft bij het bijeenbrengen van materiaal een voortrekkersrol vervuld (Liberman, 1998), later in Europa gevolgd door de European Language Resources Assocation (ELRA) en in Nederland bescheidener door het Spraak Expertise Centrum (SPEX). Het Nederlandse centrum heeft wel een belangrijke positie verworven bij de validatie van bestanden. Validatie is een harde noodzaak om de kwaliteit van bestanden te waarborgen want onbetrouwbare gegevens zijn nutteloos en belemmeren technologische vooruitgang aanzienlijk.

Ondanks de brute kracht van een HMM zijn er ook serieuze beperkingen. De eerste beperking ligt eigenlijk al voordat HMMs in het herkenningsproces gebruikt worden: in de karakterisering van het spraakgeluid. Er wordt al lang verondersteld dat er voordeel te behalen valt wanneer het spraakgeluid beschreven zou worden in analogie met de geluidverwerking van het menselijk oor. Er zijn weliswaar zeer uitgebreide modellen gemaakt van deze auditieve geluidverwerking maar als die werden gekoppeld aan systemen voor automatische spraakherkenning was er zelden sprake van een significante verbetering van de herkenningsresulaten. Uiteindelijk wordt er nu meestal alleen gebruik gemaakt van een parametrisering op een logaritmische frequentie-as. Toch verwacht men nog steeds dat door betere pre-processing winst is te behalen, in het bijzonder bij spraak in aanwezigheid van achtergrondlawaai of in de analyse van door elkaar sprekende personen.

Een werkelijke beperking van een HMM ligt in de mogelijkheid om context te modelleren. Op akoestisch niveau heeft men door de introductie van difoon en trifoon HMMs (die bijvoorbeeld de akoestische variatie in een /**a**/ beschrijven in de sequentie /**a**p/, respectievelijk /k**a**p/) een redelijke balans gevonden in de nauwkeurigheid van de beschrijving en de mogelijkheid tot schatting van de onderliggende kansen en verdelingen. Alhoewel er verderreikende invloed van articulatorisch-akoestische context bestaat dan alleen van de naastliggende spraakklanken moeten met name trifoon HMMs voldoende beschrijvingskracht bezitten voor succesvolle spraakherkenning. Wel is het modelleren van duren van spraakklanken, die soms niet en soms wel betekenisonderscheidend kunnen zijn, een blijvend probleem. Veel problematischer is echter het taalmodel waarin kansen op woordvolgordes worden neergelegd. Zelfs voor een eenvoudig bigrammodel, waarbij de kans op de opeenvolging van alle mogelijke woordparen wordt vastgelegd, is er voor een vocabulair van 10.000 woorden al een corpus van vele tientallen miljoenen woorden nodig om die kansen te schatten. Maar de syntactische en semantische context strekt zich vaak niet over twee maar over vele woorden uit. Willen we die relaties in kansen uitdrukken dan is er waarschijnlijk nog niet eens voldoende Nederlands geschreven dat als trainingsmateriaal kan dienen. Generalisaties lijken derhalve geboden.

Wie een gesproken informatiesysteem moet ontwikkelen doet er goed aan om vooraf te onderzoeken op welke manier mensen om informatie vragen. De fantasie van de onderzoeker zal blijken te kort te schieten om alle varianten zelf te bedenken, en ook zal blijken dat weinig mensen in grammaticaal correcte zinnen spreken. De conclusie moet dan zijn dat een taalmodel dat wordt ontwikkeld op basis van geschreven taal (met correcte grammatica) niet optimaal geschikt is om spontaan gesproken zinnen te beschrijven. Er kan zelfs geredeneerd worden dat alleen een beperkt aantal sleutelwoorden in een vraag om informatie van belang is en dat de inbedding in al dan niet grammaticaal correcte zinnen er minder toe doet. Deze keuze is gemaakt door Philips voor het treintijden informatiesysteem. Alleen de herkenning van sleutelwoorden is van belang en op basis van herkende sleutelwoorden leidt de computer de semantische status van de dialoog af. Sleutelwoorden zijn stationsnamen, cijfers, ‘van’, ‘naar’, ‘morgen’, ‘uur’, enzovoorts.

In de afhandeling van de dialoog moet de computer de binnengekregen informatie verifiëren en eventueel om aanvullende informatie vragen als de bedoeling van de vragensteller niet duidelijk is. Als de computer vraagt *'Hoe laat wilt u woensdag 3 juni vanuit Utrecht naar Maastricht vertrekken?'* houdt dat een impliciete verificatie in van de datum van de reis en van het vertrek- en eindstation welke gekoppeld wordt aan een verzoek om nadere informatie over het tijdstip van vertrek. Als de gebruiker geen correctie aanbrengt en alleen een tijdstip noemt gaat de computer er vanuit dat de overige informatie correct is. Als de gebruiker had gezegd *'Ik wil om negen uur naar Maassluis'* dan was de computer er vanuit gegaan dat het eindstation nu Maassluis is en was er nog een expliciete vraag gevolgd om dit te verifiëren. Door het herkennen van sleutelwoorden is de manier waarop een gebruiker reageert van minder belang. Als de gebruiker had gezegd *'Welnee sukkel, ik moet naar Maassluis, om negen uur'* had dat tot hetzelfde resultaat geleid. Op deze manier wordt uiteindelijk alle informatie verzameld en geverifieerd die nodig is voor het geven van een antwoord.

De reacties van de computer worden gemaakt met behulp van spraaksynthese. Vergeleken met automatische spraakherkenning is er op dat terrein relatief weinig gebeurd de laatste tien jaar, zij het dat voor het aanpassen van de temporele structuur van kunstmatige spraak de PSOLA techniek een grote vooruitgang bracht. Toch lijkt er nu een kentering zichtbaar te worden, misschien omdat gebruikersacceptatie van spraaktechnologie ook afhangt van een computer die verstaanbaar en natuurlijk spreekt. Daaraan lijkt het meest te worden voldaan door het aaneenrijgen van grotere stukken spraak (woorden) wanneer tenminste het aantal te spreken woorden beperkt is. Dat is in een treinreis nformatiesysteem het geval. Alleen de stationsnamen, dagen en tijden moeten in uitingen gevoegd worden en er moet een acceptable prosodie gegenereerd worden. Deze beperking geldt echter bijvoorbeeld niet voor een systeem dat e-mail berichten via de telefoon voorleest. Die tekst is onvoorspelbaar. Op dit moment is er hiervoor redelijk goede Nederlandse difoonsynthese op de markt. Een interessante ontwikkeling, die een parallel heeft met wat er op het gebied van spraakherkenning is gebeurd, is dat getracht wordt variatie in de prosodie van spraak met behulp van statistische technieken in kaart te brengen (Karaali, 1997). Tijdens de synthese kan dan gekozen worden voor de meest waarschijnlijke prosodie van een te synthetiseren zin of er kan juist wat variatie aangebracht worden door te kiezen uit goede, alternatieve patronen. Een andere brute kracht synthesemethode die karakteristiek is voor de huidige stand van de technologie is het synthetiseren van spraak door het aaneenrijgen van stukjes natuurlijke spraak die zorgvuldig uit een zeer groot spraakbestand worden gezocht. Vele uren spraak van een spreker worden opgeslagen en de computer kiest daaruit de best bruikbare stukjes om een nieuwe zin te vormen. Doordat orginele spraak wordt gebruikt is de natuurlijkheid optimaal. De kunst is natuurlijk om de juiste stukjes te vinden (Campbell, 1998).

Het is niet mogelijk om objectief vast te stellen hoe goed synthetische spraak klinkt. Daar is altijd een luisteraarsoordeel voor nodig. De verstaanbaarheid van de spraak is eenvoudig te testen door luisteraars te vragen wat ze gehoord hebben. De beoordeling van de natuurlijkheid is een stuk moeilijker. Wat dat betreft is er op het internet een interessante website (http://www.ldc.upenn.edu/ltts/) (Pols et al., 1998). Allerlei spraaksynthesesystemen, voor verschillende talen en ontwikkeld door allerlei laboratoria, zijn op die website beschikbaar om onderworpen te worden aan het vergelijkend oordeel van de bezoeker. De testzin kan door de gebruiker worden ingevoerd of er kan worden gekozen voor een zin met bepaalde eigenschappen die uit een groot tekstbestand wordt gehaald. Deze testzin kan dan

door alle relevante systemen worden uitgesproken. Op deze manier is manipulatie uitgesloten. Door de terugmelding van de gebruiker - waar ook ter wereld - kunnen de onderzoekers weer veel leren over de goede en slechte punten van de systemen.

Het testen en evalueren van automatische spraakherkenning gebeurt aan de hand van een grote hoeveelheid spraakmateriaal waarvan de tekst bekend is. Er hoeft slechts te worden vastgesteld hoeveel van die tekst door eens systeem juist wordt herkend. Evaluatie wordt ingewikkelder als het gaat om complexe systemen, zoals dialoogsystemen. Het commentaar van gebruikers kan veel achtergronden hebben en het zal niet altijd even duidelijk zijn op welke systeemcomponenten een reactie is terug te voeren. Een foutieve opmerking van de computer kan bijvoorbeeld veroorzaakt zijn door een foutieve herkenning maar ook door een foutieve interpretatie van een opmerking van de gebruiker of door een onjuiste afhandeling van de dialoog. In een test van diverse dialoogsystemen door experts (den Os en Bloothooft, 1998) bleek dat systemen beoordeeld werden op de interactieve componenten van het systeem zoals die tot uitdrukking komen in de taakvoltooing en de wijze van corrigeren van fouten, op de functionele capaciteiten van het systeem en op de kwaliteit van de computerspraak. Voor de systeemontwikkelaar is het verder van belang om te weten hoe snel een gebruiker de gewenste informatie krijgt: het Zwitserse treininformatiesysteem bleek drie keer sneller dan een Frans systeem met menusturing door toetsen. Dat geeft commercieel de doorslag.

Als een gesproken dialoogsysteem toegankelijk wordt gemaakt voor een breed publiek treden nog een aantal opmerkelijke verschijnselen op. Bij het Nederlandse treintijden informatiesysteem komt het regelmatig voor (12%; van Haaren et al., 1998) dat een gebruiker bijvoorbeeld vraagt naar een verbinding tussen Utrecht en Amsterdam waarop de computer zegt '*de trein naar Amersfoort vertrekt om negen uur tien*'. Waarop de gebruiker beleefd '*dank u wel*' antwoordt. Dit kan betekenen dat de gebruiker niet beseft dat er getracht kan worden de fout te corrigeren door te zeggen '*nee, ik wil naar Amsterdam*'. De gebruiker denkt waarschijnlijk dat aan het ongewenste antwoord niets meer te doen is. We leren hier uit dat het publiek vertrouwd moet worden gemaakt met de onvermoede mogelijkheden van nieuwe technologie. Anders dan via beeldscherm interfaces biedt een telefoonlijn weinig mogelijkheid om functionaliteiten duidelijk te maken, dat moet langs een andere weg gebeuren. Daarnaast is het ook opmerkelijk dat gebruikers beleefd zijn tegen de computer, men handhaaft de gewone en ingesleten beleefdheidsfrasen. Zeker als de computerspraak natuurlijk klinkt zal de beller vaak niet meteen beseffen met een computer van doen te hebben (als dat niet expliciet is aangekondigd). In het verlengde hiervan ligt de anekdote over de faciliteit om een telefoonverbinding te leggen louter door het noemen van een naam. Sommige mensen verbazen zich daar in het geheel niet over en vinden dat heel gewoon: natuurlijk komt je vriend(in) aan de lijn als je alleen maar '*schat*' zegt. Men heeft geen benul van de complexe technologie die dat mogelijk maakt. En misschien is dat precies zoals het zou moeten zijn. Van de meeste technologie om ons heen weten we niets.

In deze bijdrage is een gesproken dialoog systeem centraal gesteld om de ontwikkeling van spraaktechnologie aan de hand van allerlei deelsystemen te illustreren. Maar spraaktechnologie is breder en veelzijdiger. Er zou gesproken kunnen worden over het telefonisch identificeren en verifiëren van sprekers, over allerlei telefonische informatie- en besteldiensten en banktransacties, over het spreken van commando's in een auto, over toepassingen van spraakherkenning in uitspraaktraining van niet-moedertaal spekers (Sevenster et al., 1998) en in het onderwijs, over het gebruik van HMMs om akoestische

stemkarakteristieken vast te leggen (Bloothooft en Binnenpoorte, 1998) en nog veel meer. De dynamiek van het vakgebied is hopelijk echter al voldoende gedemonstreerd. Gesproken dialoogsystemen tonen daarnaast ook aan dat taaltechnologie en spraaktechnologie steeds vaker *samen* nodig zijn. En samen zullen ze naar verwachting in de informatiemaatschappij die zich aan het vormen is een centrale rol spelen. Het is een boeiend vak met een boeiende toekomst.

Referenties

Bloothooft, G. and Binnenpoorte, D. (1998). Towards a searcheable resource of phonetograms. *Proceedings Voicedata98*, Utrecht, 112-116.

Campbell, N. (1998). Design of Speech Corpora for Use in Concatenative Synthesis Systems. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 1309-1312.

Haaren, L. van, Blasband, M., Gerritsen, M., Schijndel, M. van (1998). Evaluating Quality of Spoken Dialogue Systems: Comparing a Technology-focused and a User-focuesed Approach. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 655-660.

Karaali, O. (1997). Text-to-Speech Conversion with Neural Networks: a recurrent TDNN Approach. *Proceedings Eurospeech'97*, Rhodos, pp. 561-564.

Lee, K.F., Hon, H., Hwang, M., Mahajan, S., and Reddy, R. (1989). The SPHINX speech recognition system. *Proceedings IEEE ICASSP*, pp. 445-448.

Leeuwen, D.A. van, Berg, L.-G. van den en Steeneken, H.J.M. (1995). Human benchmarks for speaker independent large vocabulary recognition performance. *Proceedings Eurospeech'95*, Madrid, pp. 1461-1464.

Liberman, M., Cieri, C. (1998). The creation, distribution and use of linguistic data: the case of the Linguistic Data Consortium. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 159-164.

Os, E.A. den en Bloothooft G. (1998). Evaluating various spoken dialogue systems with a single questionnaire: Analysis of the ELSNET Olympics. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 51-54.

Pallett, D. S. (1998). The NIST role in Automatic Speech Recognition Benchmark Tests. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 327-330.

Pols, L.C.W., Santen, J.P.H. van, Abe, M., Kahn, D., and Keller, E. (1998). The use of large text corpora for evaluating test-to-speech systems. *Proceedings 1st Language Resources and Evaluation Conference*, Granada, pp. 637-640.

Sevenster, B., Krom, G. de, Bloothooft, G. (1998). Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs. *Proceedings ESCA workshop on Speech Technology in Language Learning*, Stockholm, pp 91-94.