Stem-, Spraak- en Taalpathologie
Vol. 12, 2004, No. 3, pp. 149-159

32.8310/03/1204-149

Auditieve verwerking van spraak: Hoe weinig we ervan weten

Bert Schouten

UiL-OTS Universiteit Utrecht

De centrale vraag bij de auditieve verwerking van spraak is hoe de luisteraar erin slaagt het continu variërende en ook in andere opzichten zeer variabele akoestische signaal te herleiden tot een reeks discrete eenheden – hetzij fonemen, hetzij woorden. Enkele tientallen jaren onderzoek hebben geen bruikbaar model van dit proces opgeleverd. In het foneemperceptie-onderzoek is er nauwelijks sprake van toetsbare modellen. Zo menen we bijvoorbeeld te weten dat spraak categorisch wordt waargenomen, maar we weten niet hoe dat gebeurt – en om die vraag gaat het nu juist. Ook het woordherkenningsonderzoek heeft niet geleid tot een model van auditieve verwerking van spraak, voornamelijk omdat het zich niet met de centrale vraag bezighield.
Er bestaan goede fysiologische modellen van de verwerking van geluid, en dus ook van spraak, vanaf het perifere gehoororgaan tot in de auditieve cortex, bij proefdieren. Uit die modellen wordt duidelijk dat de herleiding van signalen tot discrete eenheden niet ergens op dit pad gebeurt: het zit dus "hogerop". De verwachting is dat onderzoek naar die hogere niveaus, dankzij de opkomst van niet-invasieve fysiologische meetmethoden die op mensen gebruikt kunnen worden, ons in de komende jaren een beter beeld zullen geven van de auditieve verwerking van spraak.

1. Inleiding

Het ligt voor de hand om een themanummer dat gewijd is aan problemen in de auditieve verwerking van spraak te openen met een artikel waarin wordt beschreven hoe de auditieve verwerking verloopt in gevallen waarin er geen problemen zijn en het verwerkingssysteem dus naar behoren functioneert. Er is per slot van rekening al zó vele decennia lang onderzoek gedaan naar de werking van het gehoor en naar de waarneming van allerlei akoestische signalen, vooral van spraaksignalen, dat de verwachting alleszins redelijk lijkt dat er inmiddels toch wel een algemeen geaccepteerd en goed getoetst model van dat proces moet zijn.

Correspondentieadres: Bert Schouten, UiL/OTS, Universiteit Utrecht, Trans 10, 3512 JK Utrecht. E-mail: bert.schouten@let.uu.nl

De praktijk is echter anders. Het enige waarover veel, zo niet alles, bekend is, is het perifere gehoororgaan. Er bestaan betrouwbare modellen van de verwerking van signalen in het binnenoor en de gehoorzenuw, en al weten we niet tot in alle elektrische en chemische details hoe de cochlea (het slakkenhuis) werkt, er is een goed werkend model van de relatie tussen de invoer via het trommelvlies en de uitvoer in de vorm van zenuwpulsen in de gehoorzenuw. Dit geldt uiteraard ook voor spraaksignalen, die automatisch dezelfde verwerking ondergaan als andere akoestische signalen. Over de representatie van signalen in de hersenstam is ook nog vrij veel bekend, vooral over de cochleaire nucleus, die het 'laagste' niveau vormt van het auditieve deel van de hersenstam, maar naarmate we verder naar 'boven' gaan, dieper de hersenen in, neemt onze kennis af. Op het niveau van de cortex kunnen we op dit moment alleen gebieden aanwijzen die bij de aanbieding van bepaalde signalen worden geactiveerd, maar wat er in die gebieden gebeurt ontgaat ons nog bijna volledig.

Deze stand van zaken is niemand te verwijten. Om processen van hogere orde te kunnen observeren heeft men niet genoeg aan psychofysische experimenten (dat zijn psychologische experimenten waarin proefpersonen moeten reageren op aangeboden stimuli), terwijl fysiologisch onderzoek naar de werking van de hersenen nog niet zo lang mogelijk is. In de komende decennia mogen we daarvan veel verwachten. Maar intussen zijn we nog steeds ver verwijderd van een antwoord op de centrale vraag voor iedereen die een model wil bouwen van de auditieve verwerking van spraak: hoe slaagt het systeem erin om uit de akoestische stroom die het oor binnenkomt te reconstrueren wat de spreker heeft gezegd?

Een definitief antwoord op die vraag is misschien principieel onmogelijk. Een luisteraar *weet* dikwijls wat een spreker gezegd heeft, ook al heeft hij maar een gedeelte echt '*verstaan*'; met andere woorden, hij heeft veel meer relevante informatie tot zijn beschikking dan in het akoestische (en visueel ondersteunde) signaal besloten ligt: taalkundige kennis, kennis van de situatie en van de spreker, kennis van de onderwerpen die worden besproken, enzovoort. Hoe die informatie gebruikt wordt verschilt waarschijnlijk van persoon tot persoon en van situatie tot situatie en laat zich daardoor moeilijk modelleren; bovendien is niet bekend hoe groot de bijdrage van deze 'top-down' informatie is ten opzichte van die van de 'bottom-up' informatie die direct uit het signaal wordt afgeleid. In dit artikel beperken we ons tot de uit het akoestische signaal verkregen informatie, in de veronderstelling dat toch minstens een deel van het signaal werkelijk verstaan moet worden voordat andere informatie een rol kan gaan spelen. Die andere informatie is bovendien niet auditief, maar cognitief, en valt daardoor buiten het bestek van dit artikel, ook al is het de vraag hoe zinvol het is cognitieve informatie buiten beschouwing te laten bij het beschrijven van de verwerking van spraaksignalen (zie Plomp, 2002).

2. Twee wegen: psychofysica en fysiologie

2.1 De psychofysische weg

2.1.1 Foneemperceptie

Heel lang, tot ruwweg 25 jaar geleden, bleef het onderzoek naar de auditieve verwerking van spraak beperkt tot de herkenning van fonemen, in de impliciete veronderstelling dat eerst de kleinste taalkundige bouwstenen moeten worden herkend, voordat de luisteraar ze aaneen kan rijgen tot grotere gehelen zoals woorden. Dat onderzoek heeft weliswaar enkele modellen opgeleverd van het spraakperceptiesysteem, maar die modellen hebben als nadeel dat ze geen concrete beschrijving leveren van de werking van dat systeem of van delen ervan. Hun toetsbaarheid is daardoor over het algemeen laag.

De eerste van die modellen is de *motor theory*, die in de vijftiger jaren opkwam, maar pas in 1965 systematisch, hoewel enigszins tendentieus, beschreven werd door een tegenstander ervan: Lane (1965). Nog weer vijf jaar later kwamen de voorstanders met een antwoord, waarin ook zij eindelijk een meer systematische weergave van hun ideeën gaven: Studdert-Kennedy, Liberman, Harris en Cooper (1970). Maar meer dan een idee bleef het niet: het idee dat de waarneming van spraak grotendeels verloopt via het productiemechanisme. De luisteraar herleidt het binnenkomende signaal tot de articulatiebewegingen die hijzelf zou moeten maken (of liever: de instructies vanuit de hersenen die hijzelf zou moeten geven) om een soortgelijk signaal voort te brengen. Bij dat idee is het gebleven: er is nooit een poging ondernomen om aan te geven hoe zo'n proces in zijn werk zou kunnen gaan; er bestaan dus ook geen voorspellingen over die getoetst zouden kunnen worden.

De belangrijkste pijler onder de motor theory is het idee van de *categorische perceptie*. Dat idee is op zich heel plausibel: als we uitgaan van de herkenning van fonemen, dan moet een luisteraar per seconde ongeveer tien fonemen kunnen herkennen. Ieder foneem telt talloze verschijningsvormen en het signaal bevat geen duidelijk gemarkeerde tijdsintervallen waarin telkens één foneem optreedt. Het is dan ook redelijk om te veronderstellen dat wij als luisteraars een zeer gespecialiseerd mechanisme hebben ontwikkeld dat ons in staat stelt gemiddeld om de 100 ms een beslissing te nemen over de foneemcategorie waartoe een klein stukje spraak behoort. De welhaast eindeloze variatie in het spraaksignaal wordt teruggebracht tot ruim 40 categorieën, afhankelijk van het aantal fonemen in een taal. Om dit idee te kunnen toetsen heeft men een specifieke methodiek ontwikkeld, op basis van het grondidee van de motor theory: als het inderdaad zo is dat de perceptie van fonemen via de productie verloopt, en als het ook nog eens zo is dat er vanuit de hersenen per foneem slechts één invariantie set articulatie-instructies wordt gegeven, dan zou men moeten verwachten dat de verschillen tussen alle varianten van hetzelfde foneem niet hoorbaar zijn. Immers: de luisteraar neemt in feite alleen de invariante articulatie-instructies waar. Die verwachting kan worden getoetst in een experiment waarin luisteraars synthetische spraakstimuli van een foneemlabel moeten voorzien ("classificatie") en waarin hun

vervolgens wordt gevraagd of ze verschil kunnen horen tussen telkens twee van deze stimuli ("discriminatie"). De voorspelling is dan dat de luisteraars alleen verschil kunnen horen tussen stimuli die zij verschillende foneemlabels hebben toegekend. Die voorspelling komt zelden uit. Al meteen de allereerste keer dat een dergelijk experiment werd uitgevoerd (Liberman, Harris, Hoffman en Griffith, 1957), moesten de onderzoekers constateren dat de resultaten niet in overeenstemming waren met hun voorspellingen. En dat is altijd zo gebleven: men vond meestal wel iets dat in de verte op een experimentele bevestiging van categorische perceptie leek, maar een echte bevestiging bleef uit, totdat Schouten en Van Hessen (1992) een geval van volledige overeenstemming tussen voorspelling en resultaten publiceerden, een overeenstemming die zij verklaarden door hun gebruik van "natuurlijke" stimuli in plaats van de synthetische stimuli die iedereen tot dan toe had gebruikt. Maar zeven jaar later moesten dezelfde onderzoekers (Van Hessen en Schouten, 1999) alweer toegeven dat er meer aan de hand moest zijn: ondanks verwoede pogingen slaagden zij er niet een tweede keer in volledig categorische perceptie aan te tonen.

Iets soortgelijks overkwam Gerrits en Schouten (2004), die lieten zien dat categorische perceptie een functie is van de discriminatietaak. Er zijn vele manieren om aan een proefpersoon te vragen of hij of zij verschil hoort tussen twee stimuli. Sommige van deze manieren dwingen de proefpersoon min of meer om de stimuli eerst te categoriseren alvorens tot een oordeel te komen; het is natuurlijk niet verrassend dat de resultaten in zo'n geval behoorlijk categorisch zullen zijn, in die zin dat verschillend gecategoriseerde stimuli veel gemakkelijker te discrimineren zijn dan stimuli die tot dezelfde foneemcategorie worden gerekend. Een voorbeeld van een discriminatietaak waarvoor dit geldt is 2IFC (two-interval forced choice), waarin de proefpersoon telkens twee verschillende stimuli krijgt voorgelegd, met de vraag wat de volgorde van die twee stimuli is. Deze vraag is in de praktijk alleen te beantwoorden wanneer de proefpersoon foneemlabels hanteert, bijvoorbeeld: de volgorde is òf /p/-/t/, òf /t/-/p/. Er zijn echter ook taken die niet of veel minder een beroep doen op het gebruik van foneemcategorieën. Gerrits en Schouten (2004) moesten zelfs constateren dat de door hen gebruikte taak (een taak met vier stimulusintervallen per aanbieding, waarbij de proefpersoon diende aan te geven of de afwijkende stimulus de tweede of de derde van de aangeboden vier was) het onmogelijk maakte foneemcategorieën te gebruiken bij de discriminatiebeslissing. Maar pogingen elders om deze resultaten te repliceren leverden weer een ander beeld op, waarbij proefpersonen duidelijk wèl gebruik maakten van foneemlabels (John Kingston, persoonlijke mededeling).

De belangrijkste conclusie lijkt te zijn: we weten niet hoe de categorische perceptie van het spraaksignaal verloopt. De experimenten die we doen hebben onvoorspelbare, wisselende uitkomsten, hetgeen betekent dat we niet alle bijdragende factoren onder controle hebben. Kortom, we weten dat het spraaksignaal wordt verwerkt tot een beperkte set discrete eenheden, maar we hebben uit onze experimenten geen enkel idee gekregen over de manier waarop dat gebeurt.

Over een tweede model kunnen we heel kort zijn: de *gestural theory* van Fowler (1986). Dit is een herformulering van de motor theory. Wederom zijn het de arti-

culatiebewegingen die worden waargenomen, al wordt hieraan toegevoegd dat dat "direct" (onmiddellijk) gebeurt. Niet duidelijk is wat deze toevoeging betekent; evenmin is duidelijk hoe het proces in zijn werk zou kunnen gaan.

Een derde idee dat heel lang grote invloed heeft gehad, is het idee van de *speech mode*: de luisteraar heeft verschillende waarnemingsmechanismen voor spraak enerzijds en voor andere soorten geluid anderzijds. Dit idee is bijna triviaal: voor iedere aangeboren en aangeleerde vaardigheid hebben we tenslotte gespecialiseerde mechanismen in ons brein. Desondanks waren er vele tegenstanders, die dus meenden dat er *geen* apart verwerkingsmechanisme is voor spraaksignalen. Een duidelijk voorbeeld van een dergelijke opstelling is te vinden bij Schouten (1980). Hoe het ook zij, tientallen jaren lang is er een gevecht geleverd tussen voor- en tegenstanders van een in wezen oncontroversieel idee. Dat gevecht heeft veel energie gekost, maar het heeft niets bijgedragen aan een oplossing van de vraag hoe het auditieve verwerkingssysteem fonemen afleidt uit het spraaksignaal.

Een vierde en laatste model is het *perceptual magnet model* (Kuhl, 1991). Dit model probeert een verklaring te vinden voor een van de meest regelmatige bevindingen uit al het onderzoek naar categorische perceptie van de laatste 45 jaar, namelijk dat luisteraars twee stimuli die tot *verschillende* foneemcategorieën behoren beter van elkaar kunnen onderscheiden dan twee stimuli die *hetzelfde* foneemlabel krijgen, ook al zijn de akoestische verschillen in beide gevallen even groot. De voor dit verschijnsel gegeven verklaring is dat de perceptieve ruimte rond het centrum van een foneemcategorie (of rond het "prototype") gekrompen is, zodat daar de stimuli perceptief dichter bij elkaar zijn komen te liggen en moeilijker uit elkaar te houden geworden zijn. Deze "verklaring" is niet veel anders dan een herformulering van het te verklaren verschijnsel zelf, maar nu onder toevoeging van een metafoor: die van een magneet die de stimuli naar zich toetrekt. Het is daarom niet verwonderlijk dat een heranalyse van de data van Kuhl (1991) door Lotto, Kluender en Holt (1998) aantoonde dat haar experiment (dat hier niet beschreven hoeft te worden) equivalent is aan de traditionele categorische-perceptie-experimenten, waarin stimuli worden geclassificeerd en gediscrimineerd.

De vier genoemde "modellen" vormen geen uitputtende opsomming, maar zijn samen representatief genoeg om de conclusie te rechtvaardigen dat 50 jaar foneemperceptie-onderzoek ons niet echt op weg heeft geholpen naar een antwoord op de vraag hoe de akoestische spraakstroom wordt geconverteerd in een discrete foneemrepresentatie. Wat we dankzij psychofysisch onderzoek wèl weten is hoe het signaal door het perifere gehoororgaan heenkomt, maar dat is het terrein van de psychoakoestiek, waar nooit naar een antwoord op onze vraag is gezocht. Een goed overzicht van dat terrein is te vinden bij Moore (2003) en in het Nederlands bij Slis (1996).

2.1.2 Woordherkenning

Het foneemherkenningsonderzoek dat hierboven is besproken gaat er, meestal impliciet, van uit dat eerst de fonemen herkend worden, voordat de luisteraar toekomt aan de herkenning van grotere gehelen zoals woorden. In het woordherkenningsonder-

zoek ligt dat anders, maar niet radicaal anders. Hoewel het idee schijnt te zijn dat veel woorden, zeker de meest frequente, als geheel worden herkend, dus zonder een fonemische tussenstap, werken de meeste modellen met een herkenningsprocedure waarin in de tijd telkens een stukje informatie wordt toegevoegd aan de over een woord tot op dat moment beschikbare informatie. Die stukjes informatie hebben ongeveer de duur van een gemiddeld foneem; soms worden ze daadwerkelijk fonemen genoemd. Wat deze modellen echter onderscheidt van de pure foneemherkenning is een sterk besef van de rol van cognitieve informatie bij de herkenning van woorden. Dit besef is al heel sterk aanwezig in het Logogenmodel van Morton (1969), en het speelt ook een grote rol in het Cohortmodel van Marslen-Wilson en Welsh (1978), en de modellen Trace (Elman en McClelland, 1985) en Shortlist (Norris, 1994).

Hoewel het onderzoek naar de herkenning van woorden veel kennis heeft opgeleverd over zaken als de organisatie van het mentale lexicon en de rol van woordfrequentie bij de herkenning, heeft het geen bijdrage geleverd aan de oplossing van onze vraag naar de auditieve verwerking van het spraaksignaal, eenvoudig omdat het altijd heeft aangenomen dat die vraag al volledig was opgelost (of elders opgelost zou moeten worden). De input voor de woordherkenningsmodellen is een reeds geheel en trefzeker geanalyseerde reeks van fonemen of soortgelijke discrete eenheden; de vraag hoe die eenheden worden afgeleid uit het signaal wordt niet aan de orde gesteld.

2.2 De fysiologische weg

2.2.1 De representatie van fonemen

Het overgrote deel van het fysiologische onderzoek aan het perifere gehoororgaan heeft altijd een directe tegenhanger gevormd van de traditionele psychoakoestiek (zie Pickles, 1988). Men probeerde een directere blik te krijgen op het gehoororgaan dan met psychoakoestische middelen mogelijk was, al moest men daarvoor uitwijken naar proefdieren. Ondanks dat laatste bezwaar is men er toch goed in geslaagd de uitkomsten van psychoakoestisch onderzoek bij mensen te doen convergeren met die van elektrofysiologisch onderzoek bij katten, woestijnratten en andere dieren, zodat met enig recht kan worden gezegd dat we een vrij compleet beeld hebben van de anatomie en de werking van het gehoor, al blijft een aantal aspecten voorlopig nog onzeker. De cochlea is een zo klein en kwetsbaar orgaan dat met name de precieze werking van bijvoorbeeld de buitenste haarcellen pas in kaart zal kunnen worden gebracht wanneer verfijndere meetmethoden beschikbaar komen.

In tegenstelling tot de psychoakoestici, hebben de fysiologen wèl onderzocht hoe spraaksignalen worden gerepresenteerd in de gehoorzenuw (die de output bevat van het perifere gehoororgaan). In het bijzonder zijn vaak klinkerrepresentaties in de geoorzenuw van katten onderzocht; een mooi voorbeeld is May (2003), die laat zien dat de volledige omhullende van het klinkerspectrum terug te vinden is in de frequentiespecifieke *discharge rate* (het aantal zenuwpulsen per seconde) van de vele vezels van die zenuw. Veel meer is er echter niet gedaan, en erg specifiek voor spraak is het niet: hetzelfde resultaat had bereikt kunnen worden met ieder willekeurig spectrum.

Duidelijk is in ieder geval dat het gehoor in staat is formantpieken en andere belangrijke spectrale kenmerken 'naar boven' door te geven.

De gehoorzenuw voert informatie toe aan de auditieve hersenstam, die uit een aantal anatomisch en functioneel onderscheidbare eenheden bestaat. Die eenheden hebben gemeen dat ze de tonotopische organisatie van het signaal (een geordende representatie van laag naar hoog van de samenstellende frequenties in het spectrum) grotendeels intact laten, maar verder doen ze nogal verschillende dingen met het signaal. Zo laat May (2003) zien dat de ventrale cochleaire nucleus (VCN), de eerste eenheid die het signaal op zijn weg ontmoet, cellen bevat (de zogenaamde *choppers*), die werken over het gehele dynamische bereik van met name spraak, en die daardoor beschouwd kunnen worden als integratoren van meerdere zenuwvezels en andere VCN-cellen die slechts een deel van dat bereik weergeven (m.a.w. alleen reageren op zachte of op luide geluiden).

De overige eenheden van de auditieve hersenstam voeren allerlei bewerkingen uit, maar tasten, voor zover bekend, de spectrale informatie van het binnengekomen signaal niet aan. Het lijkt er dus niet op dat ergens in de hersenstam enige vorm van transformatie van een continu signaal naar meer discrete eenheden plaatsvindt, al past hier enige voorzichtigheid: het onderzoek is beperkt gebleven tot synthetische, geïsoleerde en stationaire klinkers en staat dus nogal ver af van natuurlijke spraak.

Van de hersenstam komen we via de thalamus (die we hier buiten beschouwing laten) in de cortex. Het ligt voor de hand om daarbij allereerst naar de auditieve cortex te kijken, in de hoop dat daar iets gevonden kan worden dat wijst op een verwerking van het spraaksignaal in de richting van meer discrete eenheden. Meer dan enkele aanwijzingen in de vorm van plausibele maar over het algemeen nog onzekere hypothesen zijn er echter niet (Pickles, 1988). Zo lijkt het erop dat in de auditieve cortex (in bepaalde gedeelten waarvan de tonotopiciteit nog wordt gehandhaafd) complexe geluiden worden geanalyseerd, temporele discriminatie plaatsvindt, evenals absolute identificatie, en dat dit deel van de cortex zelfs een belangrijke rol speelt bij het richten van de aandacht op een bepaalde geluidsbron. Dit zijn allemaal zaken die zeer goed zouden kunnen bijdragen aan de specifieke verwerking van spraaksignalen, maar directe evidentie voor deze veronderstelling ontbreekt vrijwel, op een studie van Wong en Schreiner (2003) na, waarin de representatie van twee verschillende lettergrepen (/be/ en /pe/) in de auditieve cortex in kaart wordt gebracht, op twee verschillende tijdstippen (15 en 80 ms na het begin van de lettergreep). Er zijn verschillen te zien tussen de medeklinkers en tussen de tijdstippen in de overigens fraaie kleurenplaatjes, maar wat die verschillen betekenen is vooralsnog onduidelijk. Bovendien gaat het hier om *single-unit recordings*, met elektroden in vele aparte hersencellen, zodat er een zeer gedetailleerd beeld ontstaat, maar de prijs die hiervoor moest worden betaald is dat het onderzoek op dieren moest worden uitgevoerd. En dieren herkennen waarschijnlijk geen fonemen op een manier die iets zegt over menselijke foneemherkenning, ook niet als ze getraind worden in de herkenning van twee lettergrepen.

2.2.2 De representatie van spraak in de cortex

Op het niveau van de cortex als geheel staat het onderzoek nog in de kinderschoenen. We weten wel zo ongeveer welke gedeelten van de cortex actief zijn bij de verwerking van spraak (zie Scott en Wise, 2003), maar het is nog volstrekt onduidelijk wat daar precies gebeurt. Electrofysiologisch onderzoek in de cortex van dieren is nog veel problematischer dan het op lagere niveaus al is: het is niet te verwachten dat de representatie van voor dieren betekenisloze spraaksignalen op dat niveau te vergelijken is met die bij mensen. Veel meer kan worden verwacht van onderzoek met behulp van auditieve signalen die wèl iets betekenen voor de proefdieren; te denken valt dan vooral aan signalen van soortgenoten (zie bijvoorbeeld Esser, 2003, en Margoliash, 2003 voor onderzoek bij resp. vleermuizen en zangvogels). Dergelijk onderzoek zou ons informatie kunnen geven over de verwerking en representatie van betekenisvolle signalen in de cortex in het algemeen – hoewel uiteraard niet mag worden uitgesloten dat de menselijke verwerking en representatie van spraak misschien zoveel unieke kenmerken bevat dat onderzoek bij dieren daar weinig over kan zeggen. Het is duidelijk dat op dit niveau onderzoek moet worden gedaan bij mensen; dat kan alleen met niet-invasieve technieken zoals PET en fMRI (Scott en Wise, 2003) en ERP (Kraus en Nicol, 2003). Dergelijke technieken bestaan nog niet erg lang en ze zijn ook nog lang niet nauwkeurig genoeg in termen van temporele en spatiële resolutie om ons te helpen in onze zoektocht naar de transformatie van het auditieve signaal tot taalkundige of lexicale eenheden. Maar de verwachting lijkt gerechtvaardigd dat de voortschrijdende techniek ons in de toekomst in staat zal stellen een steeds gedetailleerder beeld te vormen van de verwerking van spraak in de cortex.

3. Conclusie

Meer dan een halve eeuw onderzoek naar de perceptie van spraak heeft ons nauwelijks nader gebracht tot een antwoord op de centrale vraag op dat terrein: hoe herleidt het waarnemingsmechanisme de continue en zeer variante stroom van het spraaksignaal tot een beperkt aantal discrete eenheden? Psychofysisch onderzoek naar foneemherkenning kon dat niet, omdat het werkte met inadequate theorieën en opliep tegen de begrensde mogelijkheden van de psychofysische methode om een zo uitermate complex proces als spraakherkenning in kaart te brengen. Psychofysisch onderzoek naar woordherkenning slaagde er evenmin in, vooral omdat het die vraag niet wenste te stellen. Fysiologisch onderzoek heeft tot nu toe ook geen greep op dit proces kunnen krijgen, vooral omdat het zich aanvankelijk beperkte tot heel simpele spraaksignalen bij dieren. Nieuwe *imaging*-technieken zijn nog niet geavanceerd genoeg voor beelden met genoeg detaillering, maar de verwachting lijkt gerechtvaardigd dat de komende decennia zeer grote verbeteringen te zien zullen geven.

We moeten tenslotte niet de illusie hebben dat we, mochten we er ooit in slagen de auditieve verwerking van het spraaksignaal te ontrafelen, dan weten hoe spraakperceptie in zijn werk gaat. Auditieve verwerking is slechts een deel van het hele proces,

dat waarschijnlijk voor iedere luisteraar anders verloopt doordat er zoveel cognitieve informatie wordt gebruikt. Zelfs signalen waarin op geen enkele wijze fonemen gecodeerd zijn worden door de meeste luisteraars zonder veel moeite 'verstaan'. Verwezen zij hier naar de zogenaamde sinewave speech, waarin de drie belangrijkste formanten in een aantal zinnen zijn vervangen door sinustonen en de stemloze gedeelten zijn vervangen door stilte (Remez, Rubin, Berns, Pardo, en Lang, 1994). Dergelijke signalen hebben in feite niets met spraak gemeen: ze bestaan uit drie pure tonen die geen enkele harmonische relatie met elkaar hebben en grotendeels onafhankelijk van elkaar in frequentie fluctueren. Ze klinken in het geheel niet menselijk en hebben geen prosodie; de intonatie die men meent te horen wordt bepaald door het verloop van de laagste toon, die niets met de grondtoon te maken heeft. Die grondtoon is afwezig en kan ook niet door het gehoor gereconstrueerd worden. Toch slaagt bijna iedereen (zelfs als hij of zij geen native speaker is) er zonder al teveel moeite in de zinnen te verstaan Het is de vraag of spraakperceptie met een zo minimale input en met een zo overweldigend aandeel aan top-down informatie ooit gemodelleerd zal kunnen worden. Voor normale, alledaagse spraakperceptie mag men er echter vanuit gaan dat de herkenning van fonemen en van woorden op basis van het akoestisch signaal een onmisbare schakel is in het hele proces, ook al zijn we er na 50 jaar nog steeds niet in geslaagd daar greep op te krijgen.

4. Summary

The central question with respect to the auditory processing of speech is how the listener manages to reduce the continuously varying and highly variable acoustic signal to a series of discrete entities, whether phonemes or words. Many years of research have not yet produced a useful model of this process. In phoneme perception research, there are hardly any testable models. We think we know, for example, that speech is perceived categorically, but we do no know how this occurs – and that is precisely the important question. Word recognition research has not yielded a model of the auditory processing of speech either, mainly because it has never been concerned about this central question.

There are good physiological models of the processing of sound, including speech, from the auditory periphery up to the auditory cortex, in laboratory animals. These models show that the reduction of the signal to discrete entities does not take place anywhere along this pathway: it must be "higher up". It is expected that the availability of non-invasive physiological scanning methods, which can be used on humans, will, in the years to come, give us a better picture of the auditory processing of speech.

5. Referenties

Elman, J.L. en McClelland, J.L. (1985). An architecture for parallel processing in speech recognition: The TRACE model. *Bibliotheca Phonetica*, *12*, 6-35.

Esser, K.-H. (2003). Modeling aspects of speech processing in bats – behavioral and neurophysiological studies. *Speech Communication*, *41*, 179-188.

Fowler, C. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics*, *14*, 3-28.

Gerrits, E. en Schouten, M.E.H. (2004). Categorical perception depends on the discrimination task. *Perception and Psychophysics*, in press, *66*, 363-376.

Hessen, A.J. van en Schouten, M.E.H. (1999). Categorical perception as a function of stimulus quality. *Phonetica*, *56*, 56-72.

Kraus, N. en Nicol, T. (2003). Aggregate neural responses to speech sounds in the central auditory system. *Speech Communication*, *41*, 35-47.

Kuhl, P.K. (1991). Human adults and infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*, 93-107.

Lane, H. (1965). The motor theory of speech perception, a critical review. *Psychological Review*, *72*, 275-309.

Liberman, A.M., Harris, K., Hoffmann, H.S. en Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358-368.

Lotto, A.J., Kluender, K.R. en Holt, L.L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America*, *103*, 3648-3655.

Margoliash, D. (2003). Offline learning and the role of autogenous speech: new suggestions from birdsong research. *Speech Communication*, *41*, 165-178.

Marslen-Wilson, W.D. en Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

May, B.J. (2003). Physiological and psychophysical assessments of the dynamic range of vowel representations in the auditory periphery. *Speech Communication*, *41*, 49-58.

Moore, B.C.J. (2003). *An introduction to the psychology of hearing*. Academic Press.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165-178.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition 52*, 189-234.

Pickles, J.O. (1988). *An introduction to the physiology of hearing*. Academic Press.

Plomp, R. (2002) *The intelligent ear*. Lawrence Erlbaum.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S. en Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129-156.

Schouten, M.E.H. (1980) The case against a speech mode of perception. *Acta Psychologica*, *44*, 71-98.

Schouten, M.E.H. en Hessen, A.J. van (1992). Modeling phoneme perception. I. Categorical perception. *Journal of the Acoustical Society of America*, *92*, 1841-1855.

Scott, S.K. en Wise, R.J.S. (2003). Functional Imaging and Language: A Critical Guide to Methodology and Analysis. *Speech Communication*, *41*, 7-22.

Slis, I. (1996). *Audiologie. Horen in een wereld van geluid*. Coutinho.
Studdert-Kennedy, M., Liberman, A.M., Harris, K. en Cooper, F.S. (1970). The motor theory of speech perception: a reply to Lane's critical review. *Psychological Review*, *77*, 234-249.
Wong, S.W. en Schreiner, C.E. (2003). Representation of CV-sounds in cat primary auditory cortex: Intensity dependence. *Speech Communication*, *41*, 93-106.