14 SPREKEN EN VERSTAAN - SPRAAKKLANKEN[1]

Louis C.W. Pols

Er zijn diverse software pakketten op de markt om teksten in te spreken in een computer in plaats van deze in te voeren via het toetsenbord. Dergelijke software is o.a. een nuttig hulpmiddel voor mensen die niet of slechts in beperkte mate in staat zijn om een toetsenbord te bedienen, zoals bijvoorbeeld RSI-patiënten. Deze toepassing vereist dat gebruikers niet alleen tekst maar ook commando's in kunnen spreken. In (1) staat wat dit op kan leveren ('|' markeert de positie van de cursor):

(1)

Gesproken invoer	*Resultaat*
"wordt wakker"	(de microfoon wordt gevoelig)
"dit is een klein testje … punt"	Dit is een fijn testje.|
"selecteer fijn"	Dit is een fijn| testje.
"klein"	Dit is een klein| testje.
"ga naar het einde van de regel"	Dit is een klein testje.|
"ga slapen"	(de microfoon wordt ongevoelig)

De opdracht *ga naar* werd door dit programma eerst niet goed begrepen, omdat het deze woorden interpreteerde als "Ghana". Hierdoor werd de opdracht niet uitgevoerd, maar het woord *Ghana* ingetypt. Pas nadat dit woord uit de beschikbare woordenschat was verwijderd, bleef de irritante verwisseling verder uit. Dit kleine voorbeeld laat zien dat het proces van automatische spraakherkenning per computer heel wat kennis vereist over spraak, het onderwerp van dit hoofdstuk.

14.1 Inleiding

Het **spraaksignaal** kan het best begrepen worden als onderdeel van de **spraakketen**: de spreker denkt ergens aan en wil iets gaan zeggen; dit leidt tot spraakgeluid dat het oor van de luisteraar bereikt; dit spraakgeluid wordt door de luisteraar gehoord en geïnterpreteerd. Dit kan weer tot een spraakuiting van de luisteraar leiden, die dan dus spreker wordt. Zo wisselen spreker en luisteraar elkaar af en blijft de spraakketen in stand (zie ook hoofdstuk 4 over beurtwisseling).

Binnen de taalwetenschap richten twee vakgebieden zich op het spraaksignaal. Het vakgebied van de **fonetiek** houdt zich bezig met het fysieke proces van spreken en verstaan en met de natuurkundige eigenschappen van het spraaksignaal: hoe worden klanken geproduceerd, wat zijn hun signaaleigenschappen, hoe worden klanken van elkaar onderscheiden in de waarneming? Het vakgebied van de fonologie (zie hoofdstukken 15 en 16) gaat over klanken als onderdeel van het taalsysteem: welke klanken zijn betekenisonderscheidend in een taal, hoe zit het klanksysteem in elkaar, op welke manier mogen klanken gecombineerd worden tot woorden en zinnen? Om deze twee verschillende perspectieven op klanken duidelijk te maken worden klanken tussen vierkante haken wanneer

[1] Dit is de *niet-finale versie* van hoofdstuk 14 in het leerboek 'Taal en Taalwetenschap' dat waarschijnlijk in de loop van 2001 zal verschijnen bij Blackwell.

het vanuit een fonetisch perspectief gaat om feitelijke klankrealisaties, en tussen schuine strepen wanneer ze als fonologische abstracties worden gerepresenteerd. Er is dus een principieel verschil tussen de realisatie van een [a] (fonetisch) en het foneem /a/ (fonologisch). Overigens zullen we de haken en strepen steeds weglaten wanneer er geen verwarring mogelijk is of wanneer dit onderscheid er niet toe doet.

In dit hoofdstuk gaat het over klanken vanuit het fonetische perspectief. Paragraaf 14.2 bespreekt de manier waarop het spraaksignaal wordt geproduceerd met behulp van de spraakorganen. In 14.3 komen de geluidseigenschappen van het spraaksignaal aan de orde. De waarneming van het spraaksignaal is onderwerp van 14.4. Vervolgens geeft paragraaf 14.5 een systematische beschrijving van de fonetische eigenschappen van spraakklanken, waarbij we ons concentreren op het Nederlands. Tenslotte komen we in 14.6 terug op de hierboven al genoemde communicatie met de computer met behulp van spraak.

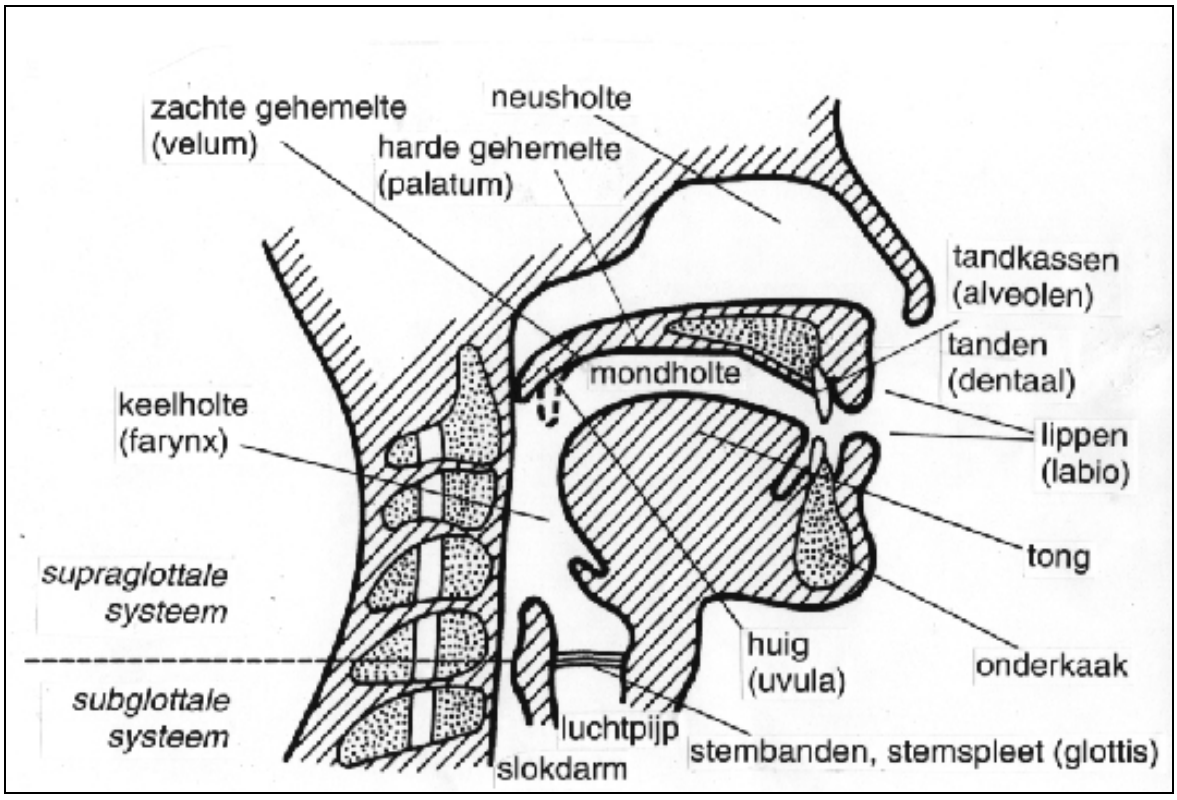
14.2 Spreken

Met de **spraakorganen** die ons ten dienste staan kunnen we allerlei geluiden maken, zoals klikken, knorren, grommen, zuchten, fluiten en brabbelen, maar we kunnen er ook de spraakklanken van het Nederlands mee produceren en we kunnen die klanken aaneenrijgen tot woorden en zinnen. We kunnen iets op een vragende, dreigende of liefelijke manier zeggen, maar we kunnen ook een hele zin fluisteren. Als we als baby in Turkije of China waren geboren en opgevoed, hadden we moeiteloos Turks of Chinees gesproken. Hoe doen mensen dat eigenlijk, spraakklanken produceren en de woorden in een taal uitspreken?

Bij het spreken fungeren de longen als een soort blaasbalg die de lucht, via de luchtpijp en de stembanden, naar het keel-, mond-, neuskanaal stuurt. Deze lucht komt uiteindelijk via de mond en/of de neus naar buiten en wordt daar hoorbaar als spraakgeluid. Door de longen daarbij harder of zachter leeg te blazen, kunnen mensen harder of zachter spreken. Ook als we ademen blazen de longen natuurlijk lucht naar buiten, maar dan staan de stembanden open en zijn er geen vernauwingen in het mondkanaal, zodat we dan vrijwel geen geluid maken. De stembanden kunnen gesloten worden en vervolgens via de lucht in trilling worden gebracht voor de zogenaamde stemhebbende klanken zoals de klinkers *ie*, *oe*, of *aa* en bepaalde medeklinkers zoals de *w* of de *l*. Als de stembanden geopend blijven maar er elders in het mondkanaal een vernauwing optreedt, kunnen we klanken maken zoals de *s* of de *p*. Als de neusholte meedoet ontstaan klanken zoals de *m* en de *n*. Figuur 1 geeft een schematische doorsnede van het hoofd met daarin de verschillende spraakorganen.

Voor die klanken waarbij de neusholte geen rol speelt, stroomt de lucht vanuit de longen via de luchtpijp door de stembanden, die al of niet in trilling kunnen komen, en dan via het mondkanaal en de lippen naar buiten. Met het achterste deel van het zachte gehemelte kan de neusholte worden geopend voor de productie van nasale klanken zoals de *n* of de *m*, of voor genasaliseerde klinkers zoals in het Franse woord *vin*.

Spreken gaat vrijwel onbewust, maar zeker niet vanzelf. We moeten plannen wat we willen gaan zeggen en als er een stoornis optreedt in de hersenen (bijvoorbeeld bij afasie, zie hoofdstuk 2), dan lukt die planning soms niet en spreken mensen vaak onverstaanbaar, ook al kunnen ze alle spreekbewegingen op zich nog wel maken. Als om wat voor reden dan ook, bijvoorbeeld bij een ernstige tumor, de stembanden niet meer functioneren, is spreken onmogelijk.



Figuur 1: Schematische dwarsdoorsnede van het menselijk hoofd met daarin aangegeven de belangrijkste spraakorganen.

Het spreken moeten we als kind aanleren en we moeten ons bovendien houden aan de conventies binnen een taal; anders zouden we onverstaanbaar zijn. Toch is er binnen die conventies ruimte voor veel variatie. Iedere spreker heeft zijn eigen spreekstijl en eigenaardigheden.

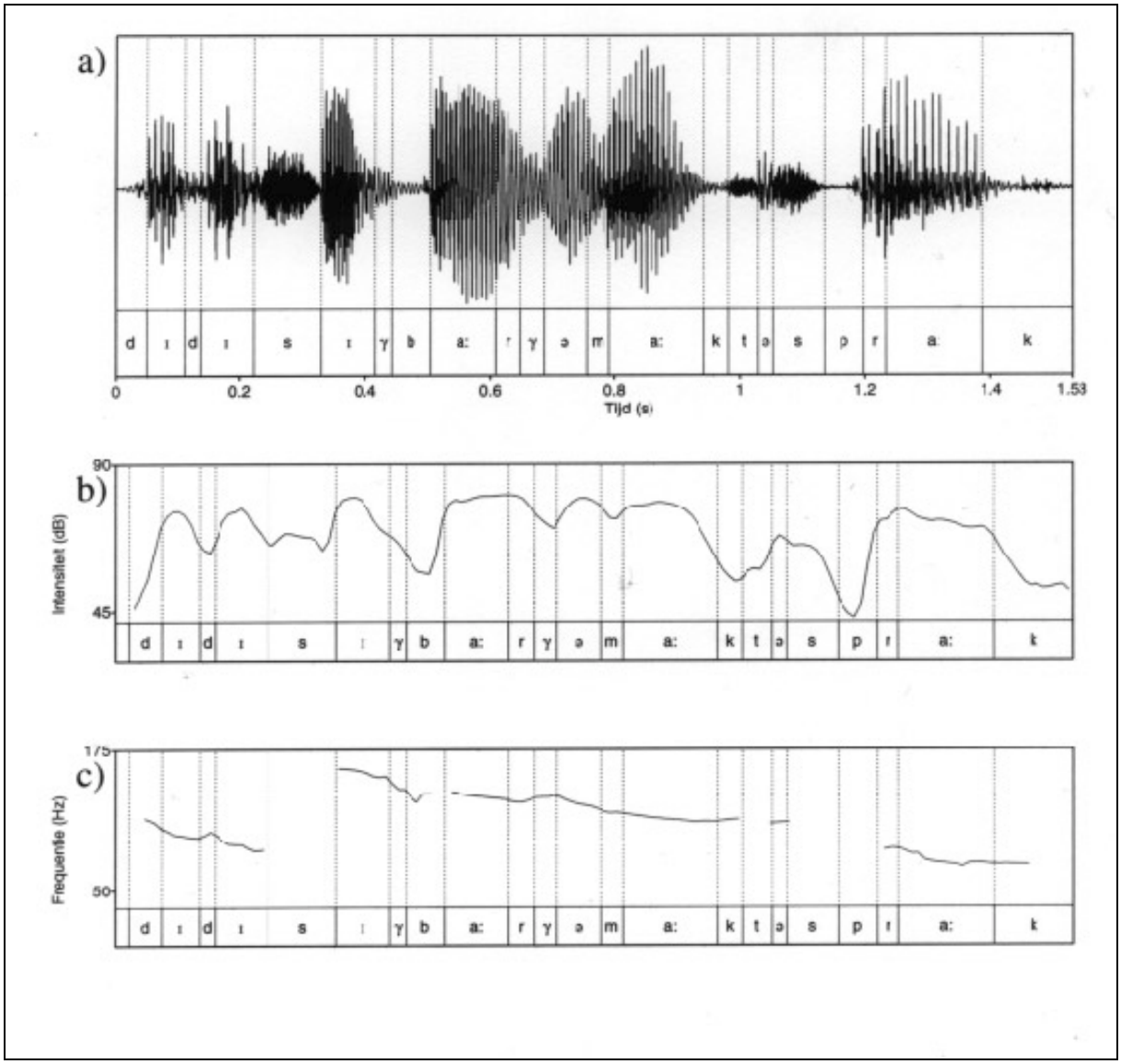
Er zijn grote verschillen tussen de gesproken taal enerzijds en handgeschreven en gedrukte teksten anderzijds. Vooral het gedrukte schrift is in principe eenduidig: iedere letter heeft zijn eigen unieke en constante vorm, woorden zijn in principe gescheiden door spaties, en zinnen zijn gestructureerd via leestekens. In de spreektaal hebben we te maken met veel meer variatie; klanken en woorden gaan meer in elkaar over en zijn niet duidelijk gescheiden van elkaar, zoals we ook in hoofdstuk 2 en hoofdstuk 11 al hebben aangegeven.

14.3 Spraakgeluid

Zonder spraakgeluid is er geen gesproken communicatie. Het spraakgeluid dat de spreker produceert, verplaatst zich meestal als trillende moleculen door de lucht. De snelheid van het geluid is niet zo verschrikkelijk groot, namelijk 340 meter per seconde, en dat is een heel stuk langzamer dan de snelheid van het licht. Vandaar ook dat je op een paar honderd meter afstand een heiblok eerder op de heipaal ziet vallen dan dat je het geluid van die klap hoort. Geluid verplaatst zich vanuit de bron in alle richtingen en moet dus een steeds groter gebied bestrijken; het wordt dan ook steeds zachter op grotere afstand van de bron.

Geluid, en dus ook spraakgeluid, kan zich echter ook via andere media verplaatsen van zender naar ontvanger, bijvoorbeeld door het water, via treinrails, of via de telefoon. Spraakgeluid is vast te leggen met behulp van een microfoon; deze zet de geluidtrillingen om in een electrisch signaal dat opgeslagen kan worden op een recorder of in het geheugen van een computer via een zogenaamde analoog-digitaal-omzetter. Dat opgenomen geluid is daarna herhaald af te spelen in zijn geheel of in kleine stukjes, waardoor het mogelijk wordt er allerlei eigenschappen van te meten.

Figuur 2 bevat een stukje spraaksignaal zoals dat door een microfoon is opgenomen. Het is het zinnetje *dit is zichtbaar gemaakte spraak*, uitgesproken door een man. Het zinsaccent ligt op de lettergreep *zicht* in het woord *zichtbaar*, als mogelijk contrast ten opzichte van een ander zinnetje over hoorbaar gemaakte spraak. Zonder al te zeer in details te treden, willen we toch enkele dingen noemen die goed aan zo'n signaal te meten zijn en die iets zeggen over het zinstempo, de klank- en woordduur, pauzes die optreden binnen en tussen woorden, de toonhoogte, de woordklemtoon en het zinsaccent.



Figuur 2: Weergave van de geluiddruk (boven), de intensiteit (midden) in decibel (dB) en de frequentie van de toonhoogte (onder) in Hertz (Hz), als functie van de tijd in seconden voor het zinnetje 'dit is zichtbaar gemaakte spraak'.

Dit zinnetje van acht lettergrepen (dit/is/zicht/baar/ge/maak/te/spraak) blijkt 1,5 seconde te duren en heeft dus een tempo van 8 : 1,5 = 5,3 lettergrepen per seconde; dat is gemiddeld 189 milliseconde (msec) per lettergreep. Het vierlettergrepige woord *eenentwintig* wordt vaak gebruikt om één seconde, oftewel 1000 milliseconden, af te passen. De gemiddelde lettergreepduur is dan dus circa 250 msec. In figuur 2 is in de vorm van een **fonetische transcriptie** ook globaal aangegeven waar zich iedere spraakklank bevindt, en dat geeft dus ook enig idee van de duur van de verschillende spraakklanken. Zo is het wel duidelijk dat de drie korte klinkers *i* allemaal een stuk korter zijn (duur circa 77 msec) dan de 3 lange klinkers *aa* (duur circa 136 msec).

Figuur 2a geeft de wisselingen in de atmosferische druk weer zoals die door een microfoon tijdens het uitspreken van het zinnetje worden gemeten. Via bepaalde berekeningen is hieruit de intensiteitscurve in decibel (dB) van deze zin af te leiden. Figuur 2b bevat deze intensiteitscurve. Hierin zijn met enige goede wil de centra van alle acht klinkers in de acht lettergrepen als maxima in de curve terug te vinden, zelfs in de korte en zwakke laatste lettergreep van *gemaakte*. Ook blijkt uit de geluidscurve in figuur 2a dat de (bijna) stille periodes in de zin niet optreden op de woordgrenzen, maar vóór de zogenaamde plofklanken *b*, *k*, en *p*. Bij dit type klanken sluit de spreker met de lippen (*p* en *b*), de tongpunt (*t*), of de tongrug (*k*) tijdelijk het mondkanaal af. Daarbij wordt enige overdruk in de mond opgebouwd, die dan met een plof ontsnapt. Zo komt de karakteristieke plofklank tot stand. Die tijdelijke afsluiting wordt zichtbaar als een korte (bijna) stille pauze in de zin, terwijl van een pauze tussen ieder woord geen sprake is. De *t* in het woordje *dit* blijkt in feite gerealiseerd te zijn als een korte *d* tussen de twee klinkers *i* in.

In figuur 2c is het zogenaamde **toonhoogteverloop** ofwel de intonatie weergegeven. Deze curve illustreert het verloop van de frequentie (Hz) waarmee de stembanden trillen gedurende dit zinnetje. Soms is de curve onderbroken; dat is in die gedeeltes van de zin waarin de stembanden niet trillen, en het signaal dan ook onregelmatig (ruizig) is, omdat hier stemloze klanken worden gerealiseerd, zoals de *s*. We kunnen zien dat de stemhebbende *z* in *zichtbaar* ook stemloos (dus als een *s*) wordt uitgesproken en samenvalt met de *s* van *dit is*. Het toonhoogteverloop is dalend. Dit is gebruikelijk voor Nederlandse bevestigende zinnen. Vraagzinnen hebben vaak juist een stijgend toonhoogteverloop. Tenslotte is aan de plaatselijke sprong in het toonhoogteverloop in de buurt van de eerste lettergreep van *zichtbaar* goed te zien dat dit woord het zinsaccent heeft.

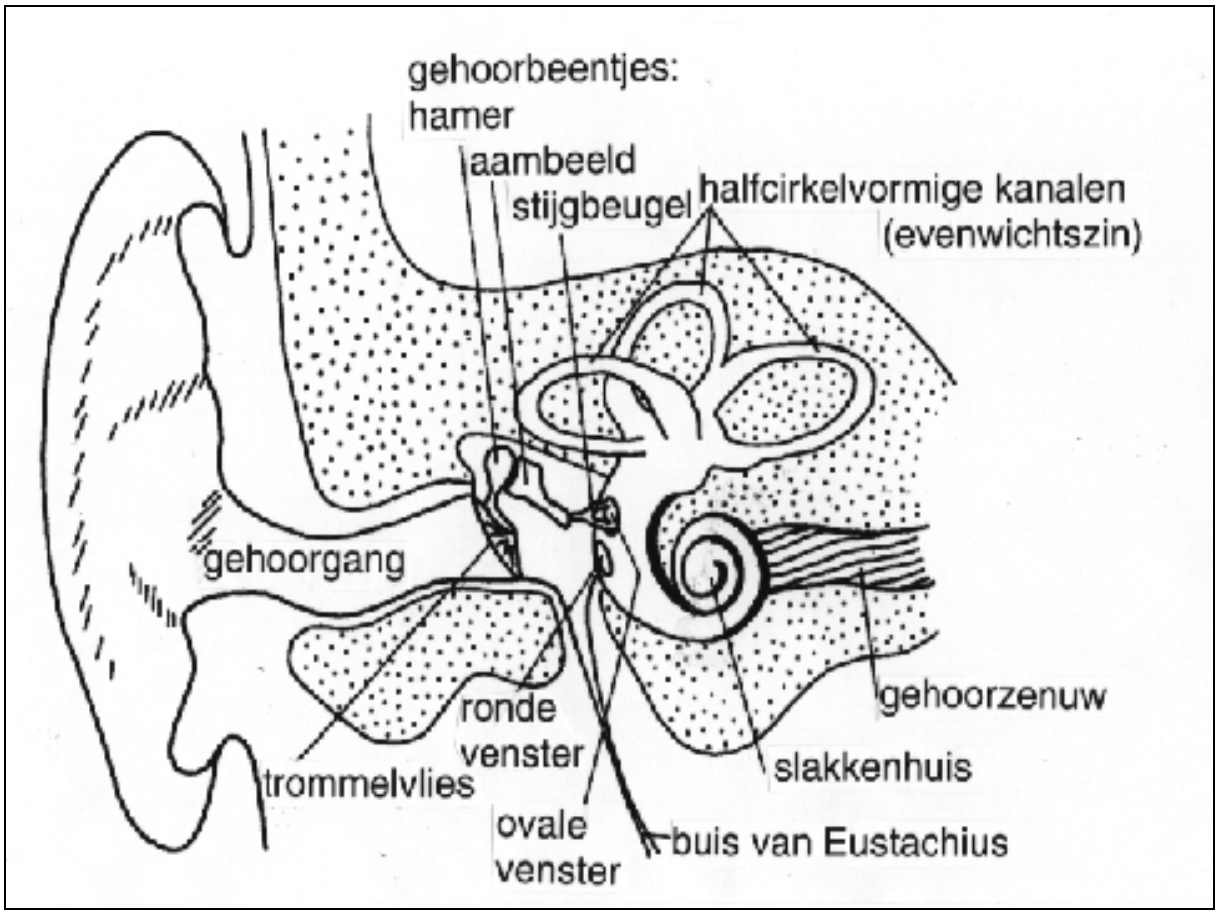
Op een aantal manieren blijkt dat in het drielettergrepige woord *gemaakte* de middelste lettergreep beklemtoond is en dat de eerste en laatste lettergreep dat minder of niet zijn. De klinker in *te* is kort, terwijl de middelste klinker *aa* lang is en luid. De eerste lettergreep van *gemaakte* is overigens ook verrassend krachtig.

14.4 Horen en verstaan

Hoe kunnen mensen eigenlijk spraak horen en verstaan? Welke processen spelen daarbij een rol? In hoofdstuk 2 is dit onderwerp al besproken vanuit het gezichtspunt van de soorten kennis die een taalgebruiker inzet. Hier komen enkele meer technische aspecten aan de orde.

Om het geluid van spraak te kunnen horen en verstaan, moeten we dat spraakgeluid

eerst via onze oren waarnemen en verwerken. Ons oor vangt de geluidtrillingen in de lucht op en voert die via de gehoorgang en het trommelvlies naar het middenoor met de drie gehoorbeentjes (hamer, aambeeld en stijgbeugel). Dit is schematisch weergegeven in figuur 3.



Figuur 3: Schematische dwarsdoorsnede van het oor met daarin aangegeven de belangrijkste gehoororganen.

De stijgbeugel brengt via het ovale venster de geluidtrillingen over naar de vloeistof in het binnenoor (slakkenhuis) waar een soort frequentieanalyse plaatsvindt. Het resultaat daarvan wordt door zenuwsignalen via de gehoorzenuw naar de hersenen gestuurd voor verdere interpretatie. Om spraak te kunnen verstaan moeten we dus allereerst goed kunnen horen, daarnaast moet er een goede verwerking in de hersenen plaatsvinden, terwijl ook de in hoofdstuk 2 besproken soorten kennis nodig zijn om het gehoorde te kunnen interpreteren. Als op een van deze drie niveaus iets mis gaat, is het spraakverstaan verstoord.

14.5 Spraakklanken

Om een taal te bestuderen of om onderzoek te doen naar een specifieke taal, of naar gesproken talen in het algemeen, is het van belang inzicht te hebben in de (structuur van de) spraakklanken van die taal. Welke klanken komen in die taal voor, hoe worden ze

geproduceerd, wat zijn hun eigenschappen en wat onderscheidt ze van elkaar in de waarneming?

Er zijn allereerst twee belangrijke categorieën van spraakklanken, namelijk die van de **vocalen** of klinkers, en die van de **consonanten** of medeklinkers. In een tweelettergrepig woord als *strepen*, zijn de lange vocaal [e:] en de neutrale vocaal schwa [ə] de kernen van de twee lettergrepen *stree* en *pen*. *Str* is een cluster van drie consonanten, terwijl ook *p* en *n* consonanten zijn. Vocalen zijn klanken die worden geproduceerd zonder vernauwing in het mond-keelkanaal. Consonanten worden juist wel met zo'n vernauwing uitgesproken.

Consonanten kunnen dan ook verder worden onderscheiden naar de mate van vernauwing die optreedt bij het uitspreken ervan. Dit heet de **manier van articulatie**. De volgende groepen van klanken verschillen van elkaar naar hun manier van articulatie:

- **plosieven** (plofklanken) , zoals de *t* en de *p*, worden geproduceerd door met behulp van een volledige afsluiting overdruk in de mond op te bouwen, om de lucht vervolgens in één keer te laten ontsnappen;
- **fricatieven** (wrijfklanken), zoals de *f* en de *s*, worden geproduceerd met behulp van een vrijwel volledige vernauwing van het mond-keelkanaal, waar de lucht als het ware doorheen geperst wordt;
- **liquidae** (vloeiklanken), zoals de *r* en de *l*, beter bekend onder de Engelse term *liquids*, worden geproduceerd door de lucht langs de zijkanten van de tong te laten ontsnappen;
- **nasalen** (neusklanken), zoals de *m* en de *n*, worden geproduceerd door de lucht door de neus te laten ontsnappen;
- **halfklinkers** (glijklanken), zoals de *w* en de *j*, worden geproduceerd met nauwelijks enige vernauwing in het mond-keelkanaal.

De *p*, *t*, en *k* zijn alledrie plofklanken of plosieven, maar toch zijn ze verschillend; dat heeft te maken met hun verschillende **plaats van articulatie**: de plaats waar een vernauwing in de mond optreedt. Voor de *p* is dat bij de lippen, voor de *t* bij de boventanden en voor de *k* bij het zachte gehemelte. Dit onderscheid in plaats van articulatie geldt niet alleen voor de plofklanken *p*, *t*, en *k*, maar ook voor andere groepen medeklinkers. De belangrijkste plaatsen van articulatie voor het onderscheiden van groepen van klanken zijn de volgende:

- labiale klanken, zoals de *b*, worden uitgesproken bij de bovenlip;
- dentale klanken, zoals de *d*, worden uitgesproken bij de boventanden;
- alveolaire klanken, zoals de *s*, worden uitgesproken bij de tandkas achter de boventanden;
- palatale klanken, zoals de *j*, worden uitgesproken bij het harde gehemelte;
- velaire klanken, zoals de *k*, worden uitgesproken bij het zachte gehemelte;
- uvulaire klanken, zoals de huig-*r*, worden uitgesproken bij de huig.
- glottale klanken, zoals de *h*, worden uitgesproken bij de stemspleet.

Dentale, alveolaire en palatale klanken vormen samen de groep van **coronale** klanken. Velaire, uvulaire en glottale klanken heten samen **dorsale** klanken. De **labiale** klanken vormen een groep op zichzelf.

Naast de manier en de plaats van articulatie is er nog een derde belangrijk onderscheid tussen de spraakklanken en dat heeft te maken met de **stemgeving**. Boven in de luchtpijp, in het strottenhoofd, zitten onze stembanden, zoals in figuur 1 te zien is. De ruimte tussen de twee stembanden noemt men stemspleet of **glottis**. Deze stembanden kunnen dicht zijn (onder andere bij het tillen van zware voorwerpen) of open staan (onder andere bij het ademhalen), maar ze kunnen ook trillen, oftewel stemgeven. Dit leidt tot **stemhebbende** klanken zoals de *b* en de *d*. Als de stembanden enigszins open staan en niet trillen kan men **stemloze** klanken, zoals de *p* en de *s*, produceren.

plaats →	labiaal		coronaal				dorsaal	
			dentaal, alveolair		palataal			
stemgeving → *manier* ↓	-stem	+stem	-stem	+stem	-stem	+stem	-stem	+stem
plofklank	**p** *p*aal	**b** *b*aal	**t** *t*aal	**d** *d*op	**tʃ** *ch*ecken(E) **tj** *tj*alk	**dʒ** *j*eep(E) **dj** *J*akarta	**k** *k*ok **ʔ** aap	**g** *g*oal(E)
fricatief	**f** *f*iets	**v** *v*ies	**s** *s*ier	**z** *z*ier	**ʃ** *sj*aal	**ʒ** rou*g*e(F)	**x** a*ch*t **h** *h*uis	**ɣ** *g*ele
nasaal		**m** *m*aar		**n** *n*aar		**ɲ** ora*nj*e		**ŋ** ri*ng*
liquida				**l** *l*euk **r** *r*euk				**ʀ** *r*euk
halfklinker		**w** *w*eek				**j** *j*euk		

Figuur 4: Het klanksysteem van de Nederlandse consonanten, uitgesplitst naar manier en plaats van articulatie, alsook naar stemgebruik (- stem of + stem).

Bij wijze van illustratie van deze verschillende onderscheidingen presenteren we nu het klanksysteem van de Nederlandse consonanten in figuur 4. Bij de weergave van de (vetgedrukte) klanken zijn zoveel mogelijk Nederlandse letters gebruikt, maar om de precieze uitspraak van bepaalde klanken te kunnen specificeren, zijn fonetische symbolen onontbeerlijk. Hoe zou je anders bijvoorbeeld kunnen aangeven dat de *g* in *gelei* [ʒ] anders klinkt dan de *g* in *gelijk* [ɣ], of dat er een verschil is tussen de Brabantse zachte *g* [ɣ] en de noordelijke stemloze *g* [x]. De International Phonetics Association (IPA) heeft de fonetische symbolen internationaal vastgelegd. Figuur 4 bevat ook voor iedere medeklinker een voorbeeldwoord, waarin de letters die corresponderen met de bedoelde klank schuin gedrukt zijn. Soms zijn de voorbeeldwoorden geleend uit het Engels (E) of het Frans (F). In dit schema is ook de glottisslag [ʔ] opgevoerd, zoals die bijvoorbeeld wordt

geproduceerd bij harde klinkerinzetten aan het begin van een woord als *aap*. Verder zijn zowel de tong-r [r] als de huig-r [ʀ] opgenomen.

Om de precieze plaats van articulatie voor de verschillende vocalen aan te geven, wordt gebruik gemaakt van twee dimensies, zowel een **voor-achter** dimensie als een **hoog-laag**, oftewel **gesloten-open** dimensie. De voor-achter dimensie heeft betrekking op de plek waar precies de tongrug een vernauwing in de mond maakt: voor in de mond, zoals bij de *ie*, in het midden, zoals bij de *aa*, of achter in de mond, zoals bij de *oe*. Met de hoog-laag dimensie wordt de mate van vernauwing beschreven in termen van de relatieve verhoging van de tongrug onder het gehemelte: een hoge tongrug, zoals bij de *ie*, een middenpositie, zoals bij de *ee*, of een lage tongrug, zoals bij de *aa*. Daarnaast spelen ook de lippen een rol: die zijn gespreid voor een *ie*, gerond voor een *oe*, en ongerond voor een *e*, zoals in *tel*. Deze karakteristieken leiden tot de weergave van de Nederlandse klinkers zoals in figuur 5. De dubbele punt in deze figuur is een indicatie voor de langere duur. De andere klinkers zijn kort, behalve dat de *ie*, *uu*, en *oe* meestal lang zijn voor een *r*. Naast de enkelvoudige klinkers heeft het Nederlands ook nog drie tweeklanken, de *au* [ɑʊ], *ui* [œʏ] en *ei* [ɛi], in woorden zoals *kou*, *huis*, en *ijs*.

plaats van vernauwing →	voor		midden			achter	
lipvorm →	gespreid	ongerond	gerond		ongerond	gerond	ongerond
mate van vernauwing ↓							
hoog	**i** k*ie*s		**y** f*uu*t			**u** t*oe*n	
midden		**ɪ** **eː** k*i*p k*ee*t	**ʏ**	**øː**	**ə** h*u*t p*eu*k d*e*	**ɔ**	**oː** t*o*p p*oo*k
laag		**ɛ** t*e*l			**aː** t*aa*k		**ɑ** t*a*k

Figuur 5: Klanksysteem van de Nederlandse vocalen, uitgesplitst naar mate en plaats van de grootste vernauwing in de mondholte en naar de vorm van de lippen.

Met al deze kennis gewapend is het mogelijk om vrij nauwkeurig aan te geven hoe woorden en zinnen (moeten) worden uitgesproken. Neem bijvoorbeeld nogmaals de zin uit figuur 2: *dit is zichtbaar gemaakte spraak*. De fonetische transcriptie van de losse woorden in deze zin zou zijn: [dɪt] [ɪs] [zɪxtbaːr] [xəmaːktə] [spraːk]. Rekening houdend met enkele hierboven gesignaleerde uitspraakwijzigingen, zou de transcriptie van de zin wellicht worden: [dɪdɪsʼɪɣbaːrɣəmaːktəspraːk]. De hier gebruikte symbolen hebben niet alleen een wetenschappelijk doel. Ook in het uitspraakonderwijs en bij het gebruik in woordenboeken kunnen fonetische symbolen nuttig zijn, bijvoorbeeld om te laten zien dat het Nederlandse woord *vergeven* uitgesproken wordt als [vərxeːvə], en niet als [vɛrxəvɛn].

Er is echter vaak een groot verschil tussen de formele uitspraak van een woord, zoals die in een uitspraakwoordenboek, en de feitelijke realisatie in vrije conversatie. Neem bijvoorbeeld een naam als *Koninklijke Marine*: het is niet ongebruikelijk om die uitgesproken te horen als [koːləkəmrinə]. Daarom staat er ergens in een van de vele teksten van Kees van Kooten en Wim de Bie ook geschreven *oppegeement*. Wie dit zo leest, weet misschien niet wat er staat. Maar wie het uitspreekt, beseft dat er ‘op een gegeven moment’ wordt bedoeld. Soms moeten we extra duidelijk articuleren om geen verwarring

te creëren, bijvoorbeeld om goed onderscheid te maken tussen de uitingen *acht alen* [ɑxt ʔa:lə] en *acht talen* [ɑxt ta:lə]. Meestal luistert het echter niet zo nauw en zal de spreker zo slordig spreken als de luisteraar hem toelaat te doen. Die hoorder zal meestal ook wel met behulp van gegevens uit de context en de situatie kunnen achterhalen wat de spreker bedoelt (zie ook hoofdstuk 2 en hoofdstuk 4).

14.6 Spraaksynthese en spraakherkenning

Hierboven is uiteengezet hoe mensen praten, hoe ze klanken maken. Hoe kan nu een computer verstaanbare spraak produceren? Een van de mogelijkheden is om de menselijke articulatie tot in alle details na te bootsen, maar dit blijkt buitengewoon ingewikkeld te zijn. Een ander uiterste is om de computer als een veredelde recorder te gebruiken. Alle mogelijke woorden en uitingen moeten dan vooraf worden opgenomen, waarna ze naar behoefte worden opgeroepen om achter elkaar te worden voortgebracht. Als de tekst of de spreekstijl steeds wisselt en het aantal woorden en zinnen erg groot wordt, is dit geen efficiënte oplossing meer. Maar voor specifieke toepassingen zoals de halteaankondigingen in bus of metro, de *voice mail* of een sprekende tijdmelder, is dit een uitstekende aanpak.

Een volgende mogelijkheid is om de te genereren spraak op te bouwen uit kleinere stukjes natuurlijke spraak, zoals enkelvoudige spraakklanken, of combinaties van twee spraakklanken, of halve lettergrepen. Vooral het gebruik van eenheden ter grootte van twee halve klanken, of **difonen**, is erg populair. Voor een woord als *extra* zijn dan de zeven volgende difonen nodig: [#ɛ], [ɛk], [ks], [st], [tr], [ra:], en [a:#], waarbij # de stilte voor en achter het woord aangeeft. De overgang van de ene klank in de andere is al in deze difoonelementen zelf opgenomen, en dit blijkt de synthesekwaliteit zeer gunstig te beïnvloeden. Voor een taal als het Nederlands, met zo'n 40 relevante klanken, zijn er dus 40 x 40 = circa 1600 difonen nodig. Maar zijn al die combinaties, dus al die difonen ook nodig? Combinaties als [a:h] of [kp] zijn toch uitgesloten in het Nederlands? Toch is het getal van 1600 difonen niet te hoog, want ook foneemcombinaties over lettergreep- en woordgrenzen heen moeten gemaakt kunnen worden, zoals bijvoorbeeld in *na hem*, of in *dakpan*. Meerdere medeklinkers achter elkaar, met als extreem voorbeeld het woord *angstschreeuw*, leveren overigens bij difoonsynthese nog wel eens problemen op.

Door simpelweg de juiste difonen achter elkaar te plakken, ontstaat er nog geen natuurlijk klinkende spraak. Daarvoor is ook nodig dat de luidheid, foneemduur en toonhoogte worden aangepast aan de gewenste prosodische structuur of melodie van de zin. Zo moet in een vraagzin de toonhoogte naar het eind van de zin oplopen en een beklemtoonde lettergreep moet luider en langer klinken dan een onbeklemtoonde.

De flexibelste, maar ook een van de moeilijkste vormen van **spraaksynthese** is die waarbij de hele klankgeneratie (dus het *voortbrengen* van de klanken) inclusief de overgangen tussen fonemen, wordt bestuurd door syntheseregels. Deze regels worden afgeleid uit gedetailleerd onderzoek naar de productie en de perceptie van natuurlijke spraak en door de regelmatigheden daarin te doorgronden.

In het voorgaande zijn we er stilzwijgend van uitgegaan dat er een fonetische transcriptie van de tekst beschikbaar is, die bijvoorbeeld aangeeft dat het woord *extra* als [ɛkstra:] moet worden uitgesproken. Evenzo moet ergens bekend zijn dat de drie *e*'s in

wegnemen, respectievelijk moeten worden uitgesproken als [ɛ], [e:], en [ə]. Dit aspect van tekstinterpretatie noemt men **grafeem-foneemconversie**: de omzetting van tekstsymbolen (grafemen) naar klanksymbolen (fonemen).

Pas als we in staat zouden zijn volledig natuurlijk klinkende synthetische spraak te produceren, zoals een verteller in staat is uit een boek voor te lezen met gebruikmaking van verschillende stemmen, spreekstijlen en emoties, dan hebben we alle aspecten van het spreken, van tekstinterpretatie tot stemrealisatie, doorgrond. Het is een uitdaging aan studenten en onderzoekers om dat te realiseren.

Veel taalkundige en technische kennis is ook vereist voor **automatische spraakherkenning**, zoals het voorbeeld aan het begin van dit hoofdstuk al duidelijk maakte. Eerst is dan een analyse nodig van het door de microfoon opgenomen spraaksignaal. De gegevens uit zo'n analyse worden vervolgens gebruikt om het patroon van alle uitgesproken woorden in de zin (in stukjes van 10 msec, of per foneem, of per woord) te vergelijken met de vooraf getrainde patronen voor alle fonemen en woorden. De beste overeenstemming leidt dan tot een hopelijk correcte herkenning. Voor bepaalde toepassingen is het voldoende als een beperkte set van losse woorden of korte commando's van een specifieke spreker correct wordt herkend. Voor andere toepassingen, zoals een informatieverstrekkend systeem, is het nodig dat een gesproken zin van een willekeurige spreker tenminste gedeeltelijk correct wordt herkend. Dat maakt het mogelijk dat de klant, via een soort vraag-en-antwoordspel, zo snel mogelijk de juiste informatie krijgt. Hierbij valt te denken aan zoiets als het via de telefoon verstrekken van informatie over de vertrektijden van treinen. De spraakherkenner moet dan tamelijk complexe zinnen als de volgende kunnen verstaan en begrijpen: 'Als ik morgen om twaalf uur in Den Helder wil zijn, welke trein moet ik dan nemen vanuit Amsterdam?'

Samenvatting

Het gebruik van gesproken taal is de meest natuurlijke vorm van communicatie tussen mensen, en in de toekomst wellicht ook tussen mens en machine. Binnen de **fonetiek** wordt het fysieke proces van spreken en verstaan bestudeerd. Daarbij wordt veel aandacht besteed aan het interactieproces tussen spreker en luisteraar (de **spraakketen**) en aan de rol die het **spraaksignaal** daarbij speelt. Bij het produceren van spraakgeluid maken sprekers gebruik van de **spraakorganen** in keel-, mond- en neusholte.

Aan een opgenomen spraaksignaal kan veel worden gemeten, bijvoorbeeld het **toonhoogteverloop** in een zin. Een **fonetische transcriptie** geeft de plaats en identiteit van iedere klank aan.

In het klanksysteem worden **vocalen** en **consonanten** onderscheiden. De klanken worden verder ingedeeld naar een aantal productiekenmerken. Voor de consonanten betreft dit de **manier van articulatie**, de **plaats van articulatie** en **stemgeving**. Wat de manier van articulatie betreft zijn vijf groepen consonanten te onderscheiden: **plosieven**, **fricatieven**, **liquidae**, **nasalen** en **halfklinkers**; wat de belangrijkste plaatsen van articulatie betreft zijn **labiale**, **coronale** en **dorsale** klanken te onderscheiden. Of een klank **stemhebbend** of **stemloos** is, wordt bepaald door het al of niet trillen van de stembanden, die zich in het strottenhoofd bevinden. De ruimte tussen de twee stembanden heet stemspleet of **glottis**.

De vocalen worden ingedeeld naar de mate en plaats van de grootste vernauwing in de mondholte, alsook naar de vorm van de lippen. Men maakt voor het specificeren van de mate van vernauwing van de **voor-achter** dimensie, en voor de plaats van vernauwing van de **hoog-laag** (oftewel **gesloten-open**) dimensie. De open lippen kunnen gespreid, gerond of ongerond zijn.

Tegenwoordig is er niet alleen aandacht voor menselijke, maar ook voor machinale spraak. Bij deze vorm van het genereren van spraak per computer (**spraaksynthese**) moet schriftelijke tekst omgezet worden in spraakklanken (**grafeem-foneemconversie**) en hoorbaar worden gemaakt. **Difonen** (combinaties van twee opeenvolgende halve fonemen) worden daarbij vaak als aan elkaar te rijgen spraakeenheden gebruikt. Naast machinale spraakproductie is er ook **automatische spraakherkenning**: het interpreteren van gesproken teksten door een computer.

Opdrachten

1 Een illustratieve manier om de spraakketen te testen is het oude kringspelletje waarbij een tekst in je oor wordt gefluisterd, die je vervolgens moet doorgeven. De hoorder moet die dan weer aan de volgende in de kring doorgeven, etc. Aan het eind wordt getest wat er over is gebleven van de oorspronkelijke uiting. Probeer dit eens uit, niet alleen met een kort zinnetje, maar ook met enkele onzinwoorden achter elkaar, zoals *let, tar, doog, goel*, of met een zinnetje in een vreemde taal. Welke aspecten van de zin zijn behouden gebleven? Hebben de deelnemers vooral geprobeerd iets betekenisvols door te geven, of zijn wellicht de klinkers beter behouden gebleven dan de rest?

2 Reïterante spraak is spraak waarbij alle lettergrepen in een zin vervangen worden door bijvoorbeeld *ma-ma ma-ma-ma*, maar alle andere signaaleigenschappen zoals pauzes, lettergreepduur, zinsritme, en dergelijke hetzelfde blijven. Probeer eens op zo'n manier de volgende zin: 'Zullen we gaan zwemmen?'. Kun je in zo'n geïmiteerde zin de woordgrenzen nog horen, de beklemtoonde lettergrepen, en een stijgende vraagintonatie?

3 *P*, *t*, *f*, en *s* zijn vier stemloze medeklinkers. Welke vier stemhebbende medeklinkers horen daarbij en waarom vormen zij paren?

4 In dit hoofdstuk hebben we het alleen maar gehad over Nederlandse klinkers en medeklinkers. Iedereen kent wel een regionaal dialect, of weet het een en ander van een vreemde taal. Produceer eens een paar spraakklanken uit dat dialect of die taal en kijk in hoeverre ze passen in het schema van de figuren 4 en 5.

5 Twee woorden die vrijwel hetzelfde worden geschreven, worden soms heel verschillend uitgesproken (*Barneveld* versus *beneveld*, *chocola* versus *cholera*, *gevel* versus *bevel*). Ook kan de betekenis van een woord verschillen met de uitspraak (bijvoorbeeld *kanon*, of *verspringen*). Bedenk zelf enkele andere voorbeelden.

6 Geef de fonetische transcriptie (zie fig. 3 en 4), volgens je eigen uitspraak, van de volgende Nederlandse woorden: hergebruik, banaan, asbak, Koninklijke Marine, en 298.

7 Onder andere via de homepage van het fonetisch instituut van de Universiteit van

Amsterdam (http://www.fon.hum.uva.nl) kun je verbinding krijgen met diverse tekst-naar-spraak synthesesystemen in diverse talen (kies daartoe eerst ‘speech-related demo’s and collections’ en daarbinnen bijvoorbeeld Bell Labs). Sommige pc’s hebben ook al voorzieningen voor spraaksynthese. Probeer eens een paar stukjes eigen tekst in het Nederlands of in een vreemde taal uit op zulke synthetisatoren, en kijk waar ze in de fout gaan.

8 Bij difoonsynthese worden klankovergangen als bouwstenen gebruikt. Hoeveel difonen zijn er nodig om het woord *verteren* te kunnen genereren, en hoeveel daarvan zijn hetzelfde?

9 Welk probleem doet zich bij grafeem-foneemconversie voor bij de twee woorden *bot* en *boten*.

Zelftoets

1 Wat wordt bedoeld met het begrip 'spraakketen'?

2 Wat is het verschil tussen de vakgebieden fonetiek en fonologie?

3 Waarin onderscheiden consonanten zich van vocalen?

4 Geef een voorbeeld van een liquida.

5 Wat is een dorsale klank?

6 Welke twee dimensies worden gebruikt om de plaats van articulatie van vocalen te beschrijven?

7 Wat wordt verstaan onder grafeem-foneemconversie?

8 Wat is een difoon?

Verantwoording en verder lezen

Voor meer informatie over de in dit hoofdstuk behandelde onderwerpen verwijzen we naar

Broecke, M.P.R. van den (red.) (1994), *Ter Sprake. Spraak als betekenisvol geluid in 36 thematische hoofdstukken*, Dordrecht: ICG Publications.

Koopmans-van Beinum, F.J. & Pols, L.C.W. (1989), *Syllabus inleiding spraakcommunicatie*, Rapport nr. 108, Instituut voor Fonetische Wetenschappen, Universiteit van Amsterdam.

Nooteboom, S.G. & Cohen, A. (1984), *Spreken en verstaan. Een nieuwe inleiding tot de experimentele fonetiek*, Assen: Van Gorcum.

Rietveld. A.C.M. & Heuven, V.J. van (1997), *Algemene fonetiek*, Bussum: Coutinho.