Geertje Dekkers

Automatische spraakherkenning niet optimaal

Psycholinguïstiek kan verbetering brengen

Nieuw in *De Academische Boekengids*: interviews met wetenschappers achter opmerkelijke onderzoeken. Deze keer: spraakherkenningscomputers kunnen moeilijk overweg met achtergrondruis en dialecten. Onderzoekster Odette Scharenborg ontwikkelde een programma dat meer recht probeert te doen aan de variatie in spraak .

Optimaal werken ze nog niet, de spraakherkenningscomputers van telefonische inlichtingendiensten ('Spreek uw woonplaats in'; 'Maastricht'; 'Maasbracht' Is dat correct?'). Vooral met achtergrondruis en dialecten kunnen ze moeilijk overweg. Onderzoekster Odette Scharenborg ontwikkelde software waarvan ze hoopt dat die daarin verbetering kan gaan brengen. Maar in de eerste plaats kan haar computermodel Fine-Tracker meer inzicht geven in de manier waarop mensen spraak herkennen. Onlangs presenteerde ze het model, dat ze aan de Radboud Universiteit Nijmegen ontwikkelde met een Veni-beurs van NWO, in de *Journal of the Acoustical Society of America* (juni 2010).

'Automatische spraakherkenning gaat nu behoorlijk goed voor duidelijk afgebakende taken', legt Scharenborg uit in haar nieuwe werkkamer in het Max Planck Instituut voor Psycholinguïstiek in Nijmegen: 'Netjes uitgesproken woorden, zonder al te veel dialect en zonder al te veel ruis, herkennen computers aardig goed. Maar daarmee lijkt het maximaal haalbare enigszins bereikt. Op de manier waarop spraakherkenning nu werkt, lijkt vooruitgang nog maar beperkt haalbaar.'

Daarom ontwikkelde Scharenborg nieuwe software, waardoor automatische spraakherkenning iets meer gaat lijken op spraakherkenning door mensen: 'Mijn onderzoek ligt op de grens tussen de psycholinguïstiek – psychologisch onderzoek naar taal – en de automatische spraakherkenning. Die vakgebieden werken weinig samen maar ze kunnen veel van elkaar leren.' Automatische spraakherkenning zou vooruit kunnen worden geholpen door te bekijken hoe menselijke hersenen met taal en spraak omgaan en de psycholinguïstiek zou computermodellen kunnen gebruiken om hypotheses te testen. Dat gaat sneller en het is goedkoper dan onderzoek met proefpersonen.

> 'DE PSYCHOLINGUÏSTIEK EN DE AUTOMATISCHE SPRAAKHERKENNING WERKEN WEINIG SAMEN MAAR ZE KUNNEN VEEL VAN ELKAAR LEREN.'

'Spraakherkenningsprogramma's knippen gesproken woorden in het algemeen in stukjes van ongeveer 10 milliseconde', legt Scharenborg uit. Een uitgesproken klank duurt al snel een paar keer zo lang. Vervolgens converteert het programma alle stukjes spraak om ze 'leesbaar' te maken voor een computer. Zo ontstaat een 'vector' met 39 getallen, waarin eigenschappen van de spraak vastliggen. Vervolgens vergelijkt de computer deze vectoren met een lexicon. Daarin staan woorden beschreven in termen van hun bijbehorende klanken. Uit een opeenvolging van vectoren leidt het programma woorden af.

Het computermodel dat Scharenborg implementeerde, Fine-Tracker, gaat nog fijnzinniger te werk. Het knipt de spraak op in stukjes van slechts 5 milliseconde. Vervolgens kent het aan eigenschappen als stemhebbendheid en gebruik van de lippen een waarde toe tussen 0 en 1.

Scharenborg: 'Letters klinken steeds iets anders, elke keer dat ze worden uitgesproken. Het zal niemand lukken de b in "boom" twee keer precies hetzelfde uit te spreken. Spraakcomputers doen alsof dat wel zo is en daardoor kunnen ze niet goed overweg met verschillende uitspraken.'
Maar spraak is dus meer dan een verzameling van steeds dezelfde klanken. Omdat Scharenborgs programma naar kenmerken zoals stemhebbendheid en de positie van de tong kijkt, in plaats van naar opeenvolgingen van klanken zoals in standaardspraakherkenners, kan het beter overweg met verschillende uitspraken van een en hetzelfde woord.

Vervolgens vergelijkt Fine-Tracker, net als andere programma's, een uitgesproken woord ('boom' bijvoorbeeld) met een grote hoeveelheid mogelijkheden uit een (reeds bestaand) lexicon – 'boom', maar ook 'boem', 'poon' enzovoort – en bepaalt het bij welk woord de uitspraak het beste past. Daarbij kan het bij Fine-Tracker gebeuren dat niet alle vectoren netjes overeenkomen met het lexicon. Maar omdat de rest wel goed past, blijkt 'boom' toch de beste match, als het goed is. Op die manier doet het programma meer recht aan de alledaagse variatie in spraak en Scharenborg hoopt dat die flexibiliteit zal helpen bij de herkenning ervan.

Het eerste positieve resultaat is er al. Uit onderzoek met proefpersonen was al bekend dat tijdsduur een belangrijke rol speelt bij spraakherkenning. Neem de woorden 'ham' en 'hamster': de lettergreep 'ham' duurt langer in het woord 'ham' dan in het woord 'hamster'. 'Hoe meer lettergrepen er in een woord volgen, hoe sneller de eerste lettergreep wordt uitgesproken', zegt Scharenborg: 'Mensen maken daar

gebruik van. Als ze de langere versie van “ham” horen, wordt in hun hersenen vooral het woord “ham” geactiveerd; bij de korte versie vooral het woord “hamster”.’ Vóórdat de uitspraak van de lettergreep ‘ham’ is afgelopen, weten mensen meestal al wat er volgt: ‘ham’ of ‘hamster’. Dat is vast te stellen in een laboratorium, waarbij mensen plaatjes van een ham en een hamster voor zich hebben. Bij de lange versie kijken ze sneller naar de ham, bij de korte sneller naar de hamster.

> ‘HET ZAL NIEMAND LUKKEN DE B IN “BOOM” TWEE KEER PRECIES HETZELFDE UIT TE SPREKEN.’

Deze kennis verwerkte Scharenborg in haar model, dat vervolgens beter functioneerde: de resultaten kwamen dichter bij die van mensen te liggen. Dat is een belangrijke stap in de goede richting. Maar, benadrukt Scharenborg, Fine-Tracker werkt nog steeds met speciaal ingesproken woorden: één stem die keurig ABN spreekt, zonder grote verschillen in volume of toonhoogte en uiteraard zonder veel ruis. ‘Het zou heel mooi zijn ooit een programma te maken dat goed zou kunnen omgaan met al dat soort verschillen’, zegt Scharenborg: ‘Idealiter zou ik één programma willen maken dat zelfs met meerdere talen overweg kan. Maar dat is nog ver weg. We gaan in de nabije toekomst mijn programma al wel testen met voorgelezen spraak: dat komt dichter bij alledaagse manieren van praten dan de spraak die voor laboratoriumonderzoek wordt gebruikt.’

Daarbij is nog veel onduidelijk. Zo heeft psycholinguïstisch onderzoek door David Gow laten zien dat Engelstaligen *ripe berries* (rijpe bessen) uitspreken als iets wat, als je het met een computer analyseert, lijkt op *right berries* (goede bessen). Toch hebben luisteraars daar geen problemen mee: ze begrijpen prima wat de spreker bedoelt. Daar is dus iets opmerkelijks aan de hand: iets wat klinkt als *right* wordt toch als *ripe* herkend. ‘Blijkbaar zit er iets in de uitspraak dat de computer niet oppikt en mensen wel, en waardoor mensen *ripe* horen en computers *right*’, zegt Scharenborg.

Misschien kunnen programma’s als Fine-Tracker daar het een en ander over duidelijk maken. Maar ook dan blijft het verschil tussen automatische spraakherkenning en de werking van menselijke hersenen groot. Scharenborg: ‘Mijn programma deelt spraak op in kleine stukjes. Dat gebeurt in onze hersenen ook, maar daarnaast herkennen we spraak nog op heel andere manieren. Er is bijvoorbeeld bewijs dat mensen woorden soms in één keer herkennen, zonder ze op te delen in stukjes. Mijn model is dus veel simpeler dan de werkelijkheid. Bovendien ontbreekt nog elke vorm van semantiek en cognitie. Die hebben ook invloed op spraakherkenning maar mijn model doet daar niets mee. En daar zijn we voorlopig ook nog niet aan toe.’