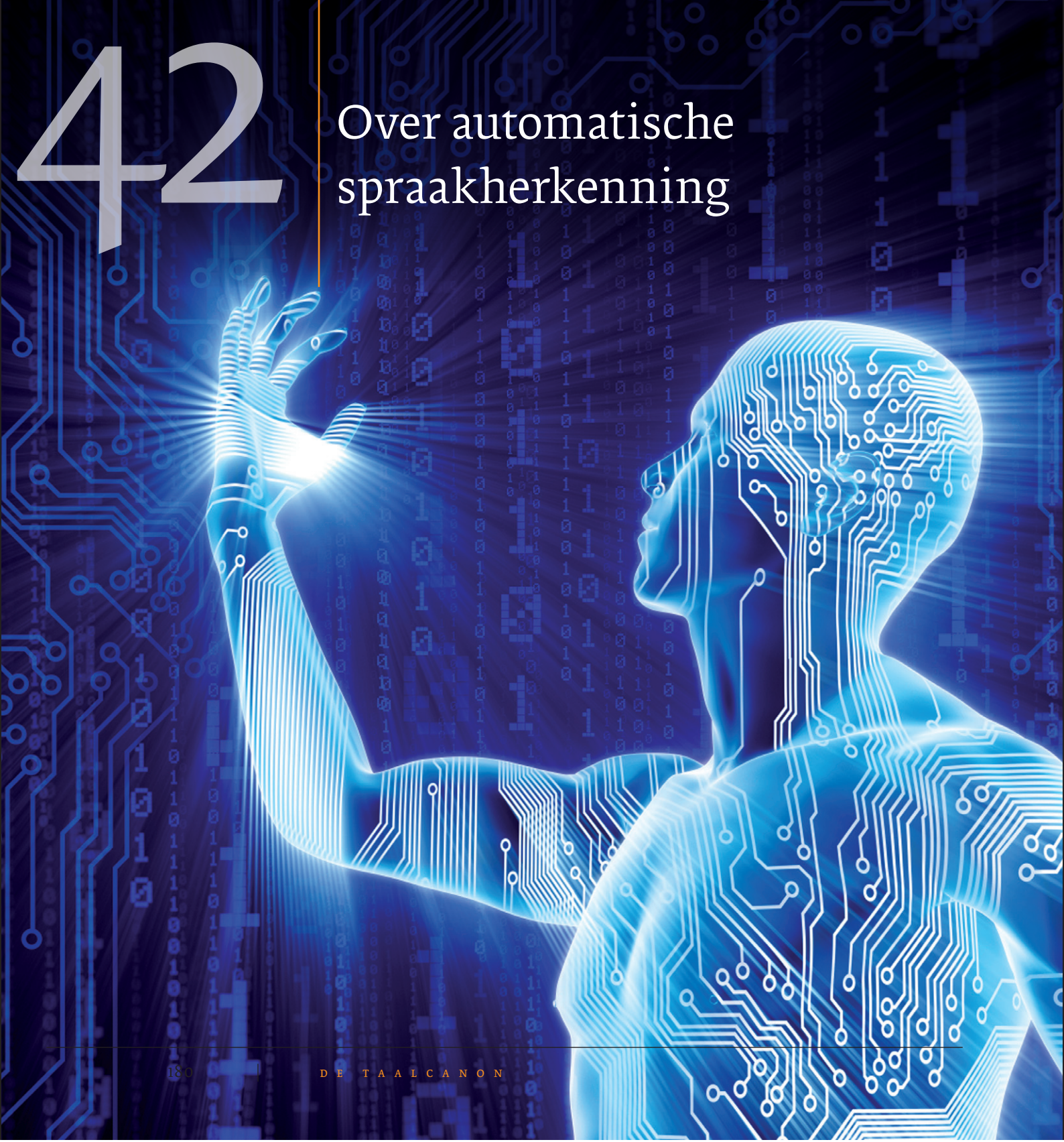
42 Over automatische spraakherkenning



Kunnen we in de toekomst praten tegen onze computer?

ARJAN VAN HESSEN

Al praat iemand met z'n mond vol, door harde muziek heen, of met een Frans accent, mensen kunnen elkaar ook in dit soort moeilijke omstandigheden vaak nog best verstaan. Maar hoe zit dat met computers? Automatische Spraakherkenning wordt al tientallen jaren gezien als een belofte die, mits eenmaal vervuld, het leven een stuk gemakkelijker zal maken. Geen lange stukken tekst meer uittikken op je toetsenbord, maar gewoon inspreken wat je op papier wilt hebben.

Dat mensen spraak kunnen ontcijferen, is niet zo vanzelfsprekend als op het eerste gezicht lijkt. Want waar woorden in geschreven taal gescheiden worden door spaties, gebruiken we in gesproken taal geen pauzes tussen de woorden die we uitspreken. Een zin als *Ik heb het formulier van de verzekeringsmaatschappij ingevuld* klinkt in gesproken taal ongeveer als: *keputformelierfandefesekringsmaatschepijingevult*. Zoals je ziet worden ook lang niet alle klanken volledig uitgesproken.

Spraakherkenning door de mens

Tijdens het luisteren bepalen we dus de woordgrenzen en herleiden we de onvolledig uitgesproken woorden tot hun officiële vorm. Dat kunnen we doordat we gebruikmaken van onze woordkennis en onze kennis over zinsbouw. Beide helpen ons om te 'voorspellen' welke woorden (en woordsoorten) we zouden kunnen horen. Bovendien zetten we onze kennis van de wereld in, of nog beter: onze kennis over het gespreksonderwerp. Stel dat we in de geluidsstroom van de voorbeeldzin zijn aangekomen bij *verzekeringsmaats*, dan weten we dat er alleen nog *chap* of *chappij* kan komen. De kans op *chap* (*verzekeringsmaatschap*) is niet zo heel erg groot omdat dat woord nu eenmaal weinig gebruikt wordt (4300 keer op Google) en dus ligt *verzekeringsmaatschappij* (175.000 keer op Google) veel meer voor de hand. Door gebruik te maken van dit soort verwachtingen, 'weten' we eigenlijk al voordat de spreker is uitgesproken welk woord er volgt. We kunnen daarop anticiperen en horen als het ware een pauze na het woord *verzekeringsmaatschappij*.

Hoe beter we een taal kennen, en hoe meer we weten over het onderwerp waarover gesproken wordt, des te beter kunnen we dus voorspellen welke woorden er zullen komen. Omgekeerd lijkt het alsof buitenlandse talen sneller klinken. Het is prettig als sprekers van zo'n taal langzaam en nadrukkelijk spreken, omdat we dan minder afhankelijk zijn van onze (zwakke) kennis van woorden en grammatica voor het decoderen (ontcijferen) van de boodschap.

Spraakherkenning door de computer

Automatische Spraakherkenning (ASR) werkt deels op

dezelfde manier als menselijke spraakherkenning: de computer verdeelt eerst het spraaksignaal in elkaar overlappende tijdsintervalletjes. Vervolgens wordt van elk tijdsinterval het spectrum berekend. Dat is de verzameling van de verschillende tonen met elk hun eigen trillingssterkte (*amplitude*). Daarna voert de computer een aantal berekeningen uit zodat het spectrum vergeleken kan worden met alle opgeslagen gegevens die horen bij de verschillende klanken. De klank die er het meest op lijkt, wordt vervolgens aan het tijdsinterval toegekend. En dan wordt het volgende intervalletje van tien milliseconden geanalyseerd, enzovoort. Voor iedere tien milliseconden is er een schatting van de klank die op dat moment werd uitgesproken. Met die opeenvolgende klanken berekent de computer vervolgens de mogelijke woorden. Zeker omdat woorden bijna nooit precies zo worden uitgesproken als officieel zou moeten, en omdat niet duidelijk is waar het ene woord eindigt en het volgende begint, is het zoeken van de woorden die bij een reeks opeenvolgende klanken horen voor een computer geen eenvoudige klus.

Vaste grammaticaherkenning Er zijn simpel gezegd twee manieren om met een computer spraak te herkennen. De eerste manier maakt gebruik van een vaste 'grammatica' waarbij de ontwerper van het programma bepaalt welke woorden op welk moment herkend kunnen worden. De tweede manier is via de 'groot vocabulaire spraakherkenning' waarmee in principe alles herkend moet kunnen worden.

Bij de' vaste grammaticaherkenning' ligt vooraf vast wat voor soort teksten mensen mogen inspreken. Dit soort systemen wordt vooral veel gebruikt wanneer duidelijk is wat de gebruiker wil. Een bekend voorbeeld is het treininformatiesysteem. Je kunt er inspreken waar je naartoe wilt en hoe laat ('*Morgenochtend om tien uur van Utrecht naar Enschede*'). Het gaat dan om een 'eindig' aantal mogelijkheden. De computer zet de ingesproken boodschap om in zogenaamde grammaticaregels. Zo'n regel is opgebouwd uit 'identifiers' (de woorden tussen vishaken):

<datum> om <tijd> van <station> naar <station>
van <station> naar <station> <datum> om <tijd>
naar <station> [vanaf|vanuit] <station> <datum>
om <tijd>

Voor de identifier <station> verwacht de spraakherkenner één van de 390 Nederlandse stations. Voor de identifier <tijd> en <datum> één van de mogelijke Nederlandse tijdsaanduidingen (8 uur 15, kwart over acht) en datumaanduidingen (morgen, volgende week maandag, tweede paasdag, etcetera).

Spraakherkenning met vaste grammatica's wordt vooral toegepast bij relatief eenvoudige, geautomatiseerde dienstverlening over de telefoon. Maar ook de nieuwste TomTom-apparaten maken er gebruik van. Een belangrijke voorwaarde is dat de gebruiker weet wat hij moet zeggen. Voor minder specifieke vragen, zoals: '*Ik wil naar de Veluwe om te wandelen*' is deze manier van spraakherkenning niet geschikt.

Groot Vocabulaire Spraakherkenning Stel, je wilt een interview met je lievelingsschrijver terugkijken in De Wereld Draait Door. Je weet alleen niet op welke dag het is uitgezonden. Op internet vind je een archief met alle uitzendingen van het afgelopen jaar. Het handigst is het als je alleen de naam van je lievelingsschrijver hoeft in te spreken

om de computer te laten zoeken. Voor zo'n zoekactie zou Groot Vocabulaire Spraakherkenning (GVSH) geschikt zijn. Bij GVSH ligt niet vooraf vast wat de gebruikers kunnen inspreken, maar moet alles dat gezegd wordt, herkend kunnen worden. De meeste spraakherkenners van dit type kunnen zo'n 64.000 verschillende woorden herkennen. De vraag is dan natuurlijk wélke 64.000 woorden, want het Nederlands heeft er veel meer.

GVSH maakt gebruik van een statistisch taalmodel. Dat is een model dat de kans berekent dat Woord-A gevolgd wordt door Woord-B (bigram), of dat Woord-A + Woord-B gevolgd worden door Woord-C (trigram). Zulke bi-, tri, quatro- en zelfs pentagrammen worden berekend met behulp van enorme hoeveelheden tekst. Zo werden aan de Universiteit Twente tien jaargangen kranten (*Volkskrant*, *NRC*, *Trouw* en *AD*) ingevoerd om de kansen van de verschillende bi- en trigrammen te berekenen. Een voorbeeld: na de woorden *ik eet* kunnen er verschillende woorden volgen, zoals *kaas, vlees, boterhammen* etc. Ook *melk* of *koffie* zouden grammaticaal correct zijn, maar de kans dat ze volgen op *ik eet* is niet heel erg groot. Helemaal onwaarschijnlijk zijn woorden als *voordeur, kerkklok* of *Klaas*. Wanneer de spraakherkenner nu na *ik eet* als volgende mogelijke woorden heeft gevonden: *Klaas, gaas, kaas en haas*, zal het statistisch model *kaas* op 1 zetten. Immers, de kans op *ik eet kaas* is vele malen groter dan *ik eet Klaas* of *ik eet gaas*.

Nadeel van het werken met deze taalmodellen is dat het nogal wat computergeheugen vraagt: met 64.000 mogelijke woorden kun je 64.000^3 = 262.144 miljard combinaties maken. Een ander nadeel is dat een taalmodel afhankelijk is van het gespreksonderwerp. Het taalmodel dat gemaakt werd met de kranten past het best bij gesprekken over het algemene nieuws, maar is minder geschikt voor spraakherkenning in een discussie tussen economen over de bankencrisis in Europa. Hoe beter een taalmodel aansluit bij het onderwerp van het gesprek, hoe beter de herkenning: in een sporttaalmodel is de kans op de woorden *Feyenoord, voetbal* en *scheidsrechter* relatief groot, in een agrarisch of politiek model is die kans vele malen kleiner.

Dicteren Terug naar de beginvraag. Is het mogelijk om een tekst te dicteren op zo'n manier dat de computer hem met zo min mogelijk fouten 'opschrijft'? Ja dat kan. Wél moet je het systeem goed trainen met je eigen stem en je beperken tot inhoudelijk gelijksoortige documenten.

Het trainen met de eigen stem is nodig om de computer te leren hoe de spreker de verschillende klanken uitspreekt. Een Tukker spreekt de o van Almelo nu eenmaal anders uit dan een Amsterdammer. Daarom krijgt de gebruiker eerst een aantal standaardteksten op het scherm die hij moet voorlezen. De computer weet wat er staat en kan dan de klankherkenning aanpassen aan de uitspraak van de spreker.

De beperking tot inhoudelijk gelijksoortige documenten is nodig om het taalmodel eenvoudig te houden. Dan werkt het beter en vlotter. Wie zowel de notulen van de hockeyclubvergaderingen wil dicteren, als rapporten over de financiële crisis, moet daarom twee profielen aanmaken. Goed getrainde sprekers die zich aan de randvoorwaarden houden, kunnen meer dan 96 procent scoren: van alle honderd uitgesproken woorden, worden er minder dan vier fout herkend.