Kunnen we in de toekomst praten tegen onze computer?

Mensen kunnen spraak onder de moeilijkste omstandigheden herkennen. 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.

Door: Arjan van Hessen

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. Een gewone zin als: "Ik heb het formulier van de verzekeringsmaatschappij ingevuld" klinkt in gesproken taal ongeveer als: "keputformelierfandefesekringsmaatschepijingevult". Wat meteen opvalt, is dat lang niet alle klanken daadwerkelijk worden uitgesproken. De officiële uitspraak van een woord als "verzekeringsmaatschappij" is "`vEr-ze:-k@-rINs-ma:t-sxA-pEi`" maar in normale spraak klinkt het meer als "f@`-se:-krINs-ma:-sxA-pEi`" of "f@`-se:-kr@ns-ma:-sx@-pEi`".

Spraakherkenning bij de mens

Tijdens het luisteren bepalen we dus de woordgrenzen en herleiden tegelijk de onvolledig uitgesproken woorden tot hun officiële vorm. Dat kunnen we door gebruik te maken van zowel de woorden die we kennen als van onze verwachting over de woorden en woordsoorten die kunnen gaan komen (grammatica). 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 hierboven gebruikte zin bij "verzekeringsmaats" aangekomen zijn. 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 waarschijnlijk volgt. We kunnen daarop anticiperen en horen als het ware een pauze na het woord verzekeringsmaatschappij.
Hoe beter we nu de taal kennen, en hoe meer we weten over het onderwerp waarover gesproken wordt, des te beter kunnen we voorspellen welke woorden er zullen komen. Het lijkt daardoor alsof een bekende taal langzamer is dan een (volledig) onbekende taal. Precies om die reden is het prettig als sprekers van een taal die we minder goed beheersen langzaam en nadrukkelijk spreken. Dan zijn we minder afhankelijk van onze (zwakke) kennis van woorden en grammatica voor het decoderen van de boodschap.

Spraakherkenning bij 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 zo'n tijdsinterval het spectrum berekend: dat is de verzameling van de verschillende tonen met elk een eigen amplitude. Voor elk spectrum berekent de computer een aantal parameters, en die worden vergeleken met alle opgeslagen parameters die horen bij de verschillende klanken. De klank die er het meest op lijkt, wordt vervolgens aan het tijdsinterval toegekend. Vervolgens wordt het volgende intervalletje van 10 milliseconden

geanalyseerd, enzovoort. Voor iedere 10 milliseconden is er dan een schatting van de klank die op dat moment werd uitgesproken. Met die opeenvolgende klanken berekent de computer dan vervolgens de mogelijke woorden. Zeker omdat we weten dat woorden bijna nooit zo worden uitgesproken zoals dat officieel zou moeten, is het zoeken van de woorden die bij een reeks opeenvolgende klanken horen geen sinecure. Bovendien moet de computer rekening houden met het feit dat we in normale spraak geen pauzes gebruiken tussen de woorden en dat je dus zonder het te merken van het ene naar het volgende woord gaat.

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 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 gesproken input mensen mogen geven. Dit soort systemen wordt vooral veel gebruikt wanneer duidelijk is wat de gebruiker wil. Een bekend voorbeeld is het treininformatiesysteem. Je kunt er inspreken van waar naar waar je wilt reizen, wanneer en hoe laat ("morgenochtend om 10 uur van Utrecht naar Enschede"). Daarbij is het aantal opties beperkt. 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 dan een van de 390 Nederlandse stations. Voor de identifier <tijd> en <datum> een van de mogelijke Nederlandse tijdsaanduidingen (8 uur 15, kwart over acht) en datumaanduidingen (morgen, volgende week maandag, 2de paasdag, etc.). 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/zij 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 DWDD. Je weet alleen niet op welke dag het is uitgezonden. Op internet vind je een archief met alle uitzendingen van het afgelopen jaar. Idealiter zou je de naam van je lievelingsschrijver intypen in een zoekveld, en de computer laten zoeken naar het juiste fragment in de juiste aflevering. Voor zo'n zoekactie zou Groot Vocabulaire Spraakherkenning (GVSh) geschikt zijn. Bij GVSh is er geen ontwerper die bepaalt hoe gebruikers moeten spreken, en in principe moet alles dat gezegd wordt, herkend kunnen worden. De meeste spraakherkenners van dit type kunnen zo'n 64.000 verschillende woorden herkennen, maar de vraag is natuurlijk wélke 64.000 woorden, want het Nederlands kent veel meer woorden.
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). Deze 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 herkenner nu (na de woorden “ik eet”) de volgende mogelijke woorden heeft herkend (Klaas, gaas, kaas, haas) dan zal het statistisch model het derde woord (=kaas) toch op 1 zetten. Immers, de kans op “ik eet kaas” is vele malen groter dan “ik eet Klaas” of “ik eet gaas”.
Nadeel van deze manier van herkennen is dat je relatief zware computers nodig hebt voor de taalmodellen. Met 64.000 mogelijke woorden kun je 64.000^3 = 262144 miljard combinaties maken. Een ander nadeel is dat zo’n taalmodel afhankelijk is van het gespreksonderwerp. Het taalmodel dat gemaakt werd met de kranten past het best bij gesprekken over het algemene nieuws. Voor het herkennen van een gesprek over de situatie van de banken in Europa voldoet het stukken minder: daarvoor zou je juist het Financieel Dagblad moeten gebruiken. 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 hoog, terwijl dat in een agrarisch of politiek model juist relatief laag is.

Dicteren

Terug naar de beginvraag. Is het mogelijk om een tekst te dicteren zodat de computer deze 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 krijg 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% scoren: van alle honderd uitgesproken woorden, worden er minder dan vier fout herkend.