

Bij mobiele telefonie en automatische informatiediensten wordt vaak gebruikgemaakt van geavanceerde spraaktechnologie: de automatische spraakherkenning. Waarom is de techniek van de spraakherkenning zo gecompliceerd? En hoe begrijpt de computer wat we tegen hem zeggen? Over het hoe en wat van de moderne spraakherkenner.

‘Dat heb ik helemaal niet gezegd!’
De prestaties van de spraakherkenner

Helmer Strik

Automatische spraakherkenning lijkt misschien iets van de laatste tien jaar, maar dat is niet zo. Al rond 1870 probeerde Alexander Graham Bell een apparaat te ontwikkelen dat spraak kon omzetten in tekst. Zijn vrouw was namelijk doof, en voor haar en andere mensen met een gehoorstoornis probeerde hij spraak zichtbaar te maken. Dat lukte hem helaas niet, maar hij slaagde er in 1874 wél in om een techniek te ontwikkelen waarmee het geluidsspectrum opgesplitst kan worden. En en passant vond hij de telefoon uit.

Spraakherkenningspakketten

Pas zo’n tachtig jaar later, in 1952, werd op de befaamde Bell Laboratories de eerste spraakherkenner gebouwd. Hierbij werd gebruikgemaakt van Bells techniek om het geluidsspectrum op te splitsen. Deze spraakherkenner kon tien uitgesproken cijfers herkennen, en deed dat in ongeveer 97% van de gevallen correct.

\$In de vijftig jaar erna zijn spraakherkenners langzamerhand steeds beter geworden. Dat kwam doordat er betere technieken en programma’s ontwikkeld werden, maar ook voor een groot deel doordat de hoeveelheid geheugen en de rekencapaciteit van computers steeds groter geworden zijn. Wie thuis een flinke pc heeft, kan daarop nu ook zelf met spraakherkenning werken. Daarvoor koop je in de pc-winkel voor een paar honderd gulden een van de bekende spraakherkenningspakketten, die installeer je op je pc, je traint het door een stuk tekst voor te lezen, en het programma is klaar voor gebruik. Je kunt dan dicteren (tekst inspreken in plaats van intypen) en je pc besturen met spraak (‘command and control’).

\$Bij deze toepassingen spreek je in een microfoon die verbonden is met je eigen pc. Er zijn ook toepassingen waarbij je de telefoon als microfoon gebruikt, en dan vindt de spraakherkenning plaats op een computer die ergens anders staat. Op dit moment kun je op deze manier informatie opvragen over bijvoorbeeld beurskoersen of treinreizen.

Informatiebronnen

Hoe krijgt een spraakherkenner het voor elkaar om spraak om te zetten in woorden? Daarvoor gebruikt hij drie informatiebronnen, die mensen ook gebruiken: akoestische modellen van de klanken die voorkomen in een taal, een woordenlijst en een taalmodel.

\$Akoestische modellen bevatten informatie over de variatie in de afzonderlijke klanken. Op basis van zo’n model maakt de spraakherkenner een eerste keuze voor de klanken die hij waarneemt. Het verschil tussen een *eu* en een *ui* wordt in eerste instantie alleen opgevat als een klankverschil. De woordenlijst bevat alle woorden die de computer moet kunnen herkennen. Die woorden staan in twee versies genoteerd: in de officiële schrijfwijze (de spelling in zogeheten grafemen), en op een aantal manieren waarop ze uitgesproken kunnen worden (de notatie in ‘fonemen’). Op basis van de woordenlijst weet de computer bijvoorbeeld dat hij het woord *duir* niet gehoord kan hebben (want dat bestaat niet), maar *deur* wel. Het taalmodel, ten slotte, bevat informatie over de waarschijnlijkheid van het vóórkomen van woorden: hoe vaak ze worden gebruikt, welke woorden vaak naast elkaar staan, enzovoort. Op basis van het taalmodel kan de computer bepalen dat het woord *huis* in de zin ‘In Holland staat een ...’ waarschijnlijker is dan bijvoorbeeld *heus*.

\$Op het eerste gezicht lijkt spraakherkenning zo een eenvoudig proces. Je herkent de afzonderlijke klanken, en kijkt of die gegroepeerd kunnen worden tot klankvormen van bestaande woorden. Bovendien heb je ook nog eens de hulp van een waarschijnlijkheidsmodel, waarmee je veel sneller kunt bepalen welke woorden er gesproken worden. De indruk dat spraakherkenning een gemakkelijke taak is, kan ook gemakkelijk ontstaan omdat het bij de mens allemaal vanzelf gebeurt. Niemand denkt bewust na bij het onderscheiden van woorden en klanken. Spraakherkenning is voor een computer echter veel minder makkelijk dan je zou denken. Met welke problemen kan een computer bijvoorbeeld te kampen krijgen?

Problemen

Een van de eerste moeilijkheden voor een goede herkenning is het probleem van de uitspraakvariatie (zie het kader 'Praten tegen de NS'). We slikken vaak veel klanken in, en zeggen dan bijvoorbeeld [amsedam] in plaats van [amsterdam], of [heboe] in plaats van [heleboel]. Soms voegen we ook klanken toe: [Delleft] in plaats van [Delft]. Dat is niet zozeer een kwestie van slordigheid, als wel van normale uitspraak. Hoe dan ook, in al deze gevallen moet de spraakherkenner weten dat het niet om verschillende woorden gaat, maar om uitspraakvarianten van hetzelfde woord. Het is praktisch onmogelijk om al die uitspraakvarianten in een woordenlijst op te nemen.

\$In de tweede plaats spreken we lang niet altijd vloeiend. Integendeel (zie hetzelfde kader). In onvoorbereide spraak zitten veel pauzes (soms stiltes, die soms worden gevuld met een aangehouden *uh*), we maken woorden half af, herhalen (stukken van) woorden, stotteren, en nog veel meer. Probeer de volgende uiting maar eens hardop aan iemand voor te lezen, en kijk eens hoe lang het duurt voordat die verstaan wordt:

maardurzij-n ookdinguwaarf- ik- n- fin ... nietfint ... tatsukloppu

Een derde probleem is de woordvolgorde. Je zou geneigd zijn om als taalmodel voor de spraakherkenner een grammatica van het Nederlands te gebruiken, een grammatica die precies uitdrukt welke woorden er in welke volgorde in een uiting mogen staan. Dat is geprobeerd, maar het werkt totaal niet. We spreken namelijk vaak niet-grammaticale zinnen uit (zie nogmaals het kader 'Praten tegen de NS'), en die zijn dan dus ook vaak niet vloeiend. Een spraakherkenner moet bijvoorbeeld nog maar begrijpen dat in *van Amste- uh Nijmegen* het woordje *van* bij *Nijmegen* hoort en niet bij het onafgemaakte *Amsterdam*.

\$Deze drie problemen hebben alleen nog maar te maken met het herkennen van de voortgebrachte spraak. Maar daarnaast vangt de microfoon vaak ook andere geluiden op. Dit kunnen allerlei achtergrondgeluiden zijn: spraak van iemand anders in de kamer, de radio, muziek, verkeerslawaai of deuren die worden dichtgeslagen. Al deze stoorsignalen kunnen het herkenproces bemoeilijken, zowel bij de machine als bij de mens.

Wenkbrauwen

Naast de drie hierboven genoemde informatiebronnen (akoestisch model, woordenlijst en taalmodel) gebruiken menselijke luisteraars bij het verwerken van spraak ook vaak visuele informatie. We kijken naar de lippen, de wenkbrauwen, de gehele mimiek en de gebaren die gemaakt worden. Verder gebruiken we in gesprekken veel zogeheten back-channel-geluiden. Dit zijn (meestal korte) geluiden waarmee we duidelijk maken of we het wel of niet met onze gesprekspartner eens zijn.

\$En ten slotte maken mensen bij hun communicatie natuurlijk op ruime schaal gebruik van hun kennis van de wereld en de context van het gesprek. Gisteren hoorde ik op tv de zin "Vandaag is Wall Street ingestort." Om dit te verstaan, gebruikte ik de context en mijn wereldkennis. Zonder die twee informatiebronnen had ik mogelijk heel wat anders verstaan. Maar het gaat ook weleens mis. Vorige week zei ik in een college

fonetiek tegen mijn studenten: “Als ik bijvoorbeeld het woord *R2D2* zeg, dan ...”. Bijna niemand had verstaan wat ik gezegd had. Sommigen kennen *R2D2* helemaal niet (wereldkennis), en anderen verwachten het niet op een college fonetiek (context). In de context van een gesprek over de *Star Wars*-films, waarin een robot met deze naam voorkomt, hadden veel meer van deze luisteraars het correct verstaan. Al deze extra informatiebronnen worden door de huidige spraakherkenners niet of nauwelijks gebruikt. Mede daarom herkent de mens spraak nog steeds beter dan een spraakherkenner dat doet. Er wordt al wel onderzoek gedaan naar het gebruik van visuele informatie voor spraakherkenning. Ofschoon het nog niet meevalt om uit een videobeeld informatie over lippen, mimiek of gebaren te halen, kan met deze extra informatie het herkenproces wel verbeterd worden.

Multimodaal

Wat is nu de toekomst van automatische spraakherkenning? Heel lang is beweerd dat spraak de meest natuurlijke manier van communiceren met een computer zou zijn, en dat spraakherkenning daarom zo belangrijk is. Daar komt men nu wel een beetje van terug, en terecht. We gebruiken immers ook andere kanalen (zoals visuele informatie), en voor communicatie tussen mens en machine is het bovendien vaak makkelijker om iets op een scherm te zien of op een scherm aan te wijzen dan om alles weer te geven met behulp van spraak. We zien dan ook dat de nadruk die in het verleden vooral lag op monomodale communicatie (alleen spraak), langzaam aan het verschuiven is naar multimodale communicatie. De invoer is dan niet alleen maar spraak, maar ook bijvoorbeeld visuele informatie (lippen, mimiek, gebaren) en informatie over de plaats waar het beeldscherm aangeraakt is (de bekende ‘touch-screens’, aanraakschermen, die al in veel toepassingen gebruikt worden).

Een paar weken geleden was ik met mijn zoon in Legoland in Denemarken en daar stonden veel van deze touch screens, die zeer goed werkten. In die herrie was gesproken communicatie zeker minder effectief geweest. Enerzijds door het lawaai op de achtergrond, en anderzijds omdat we geen Nederlands konden gebruiken. Wij spreken geen Deens, dus we zouden waarschijnlijk in het Engels hebben moeten communiceren met die informatiezuil, en het is bekend dat het herkennen van niet-moedertaalsprekers een stuk minder goed gaat. Hoewel ik zelf aan automatische spraakherkenners werk, was ik blij dat er daar geen in gebruik was.

Automatische spraakherkenning is dus verre van een opgelost probleem. Hoewel er elke dag vorderingen worden gemaakt, zijn we meer dan honderd jaar na A.G. Bell, en meer dan vijftig jaar na de eerste spraakherkenner van Bell Labs nog ver verwijderd van een automatische spraakherkenner die de prestaties van de mens evenaart. Maar als er goed nagedacht wordt over waar en hoe je een automatische spraakherkenner met zijn beperkte prestaties moet inzetten (dus niet in Legoland bijvoorbeeld), dan kan het ook nu al een nuttig instrument zijn in de dagelijkse communicatie. En daar was het allemaal toch maar om begonnen.

XXXXXXXXXXXXXXXXXXXX kader 1 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Zeven tips om een spraakherkenner op de kast te krijgen

Het kost geen enkele moeite om een spraakherkenner, vooral die in beperkte informatiesystemen, slecht te laten presteren. Als je dat wilt, bijvoorbeeld in een kritische bui of om grappig te zijn, volg dan de volgende aanwijzingen op. Succes verzekerd. (Het spreekt vanzelf dat je om hem te laten schitteren precies het tegenovergestelde moet doen.)

Spreek met afwijkende snelheid: veel te snel praten of juist veel te langzaam zet de spraakherkenner snel op het verkeerde been.

Gebruik vreemde uitspraakvariaties: slik klanken in, of verzin ze erbij.

Stotter, hikkel, haper en verbeter jezelf voortdurend.

Forceer je stem: lekker hard schreeuwen helpt. En het lucht op.

Spreek in een lawaaierige omgeving. Hoe meer achtergrondgeluid, hoe beter.

Gebruik een afwijkende woordvolgorde: een treininformatiesysteem zal bijvoorbeeld van Amsterdam naar Alkmaar verwachten. Zeg bijvoorbeeld tot Alkmaar, nu in Amsterdam.

Dwaal af: overbodige toevoegingen die niets met het gesprek te maken hebben, brengen een gesproken informatiesysteem altijd in verwarring. Daar krijg je trouwens een menselijke gesprekspartner ook gemakkelijk mee in de gordijnen.

xxxxxxxxxxxxxxxxx einde kader 1 xxxxxxxxxxxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxxxxxxxxx kader 2 xxxxxxxxxxxxxxxxxxxxxxxxxxx

Uitspraakvariatie

Het volgende voorbeeldje uit een opname van normale spraak illustreert een typische uitspraakvariatie. De transcriptie staat omgekeerd afgedrukt:

... benikbangdatwuobugeefmentonsteetsferdurbugeevu ...

[anders afdrukken:]

... ben ik bang dat we op een gegeven moment ons steeds verder begeven ...

xxxxxxxxxxxxxxxxxxxxxxxxx einde kader 2 xxxxxxxxxxxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxxxxxxxxx kader 3 xxxxxxxxxxxxxxxxxxxxxxxxxxx

Praten tegen de NS

Dit zijn een paar fragmenten uit opnamen van mensen die tegen het OV-Reisinformatiesysteem van de NS spraken. Dat gaat niet altijd even vloeiend:

- *nnn nnn ehm ehm ehm delft*
- *zandvoort vanuit zandvoort nnn ijmuiden toe*
- *eh nee ik eh dank u zeer eh stem dag stem*

En ook niet grammaticaal:

- *nnn even over nadenken vijftien uur*
- *eh ik wil morgen naar van maastricht naar roermond*
- *is wel goed morgen*
- *eh kwart over twaalf oh nee doe maar tien voor half een*
- *zeven uur negentien honderd uur wil ik vertrekken*
- *dertien uur dertig is het nu al oh kijk eens aan eh naar dordrecht*

Hier en daar klinkt ook wel emotie:

- *den bosch begrijpt vandaag hij begrijpt den bosch niet*
- *ohoho amsterdam naar assen ik heb dat dus niet gezegd he neen*
- *god nou weet ik het zelf ook niet meer*
- *ja he dat is mooi zeg doet ie goed*
- *nnn dat heb ik helemaal niet gezegd nnn*
- *nou klopt er helemaal niets meer van*
- *ik geloof dat je een beetje in de war bent nou*

Soms geheel buiten de context:

- *ja graag doe dat maar eens eventjes*
- *nee dat heb ik net gezegd maar ik bedoel*

In combinatie is het allemaal nog erger:

- *holten ik wil naar gorinchem eh ik wil op eh waar wil ik op eh op maandag*
- *eh ik wilde reizen eh morgen van haarlem naar arnhem en in arnhem aankomen*
- *omstreeks elf uur en graag met de sneltrein*

Het is eigenlijk verbazingwekkend dat het nog zo vaak goed gaat ...

xxxxxxxxxxxxxxxxxxxxx einde kader 3 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx