Call!

PRATEN IS HET NIEUWE TIKKEN

Volgens marktonderzoeksbureau comScore zal er in 2020 meer gebruikgemaakt worden van spraak dan van het toetsenbord. Hoog tijd dus om de nieuwste ontwikkelingen in spraaktechnologie op een rij te zetten.

Rob Feenstra
Projectleider/consultant bij de Universitaire Bibliotheken Leiden; heeft als aandachtsgebieden bibliotheeksystemen en de digitale bibliotheek

'Het is niet voor niets dat Amazon, als grootste internetverkoper, ook marktleider is op het gebied van spraakgestuurde apparaten'

Ergens aan het begin van deze eeuw maakte ik kennis met spraakherkenning. Een collega met RSI kreeg spraaksoftware op zijn pc. Wanneer zijn deur openstond, kon je hem al van verre met een toenemende wanhoop in zijn stem 'Enter, ENTER' horen roepen, in de doorgaans ijdele hoop dat zijn smeekbede zou worden omgezet in een bruikbaar commando.
In die tijd waren dergelijke systemen gebaseerd op *template matching*, een techniek waarbij geluid werd omgezet naar cijfercodes die vervolgens werden opgeslagen. De software werd getriggerd wanneer op een later moment iemand eenzelfde geluid liet horen. Het was daarbij noodzakelijk om langzaam en duidelijk te spreken in een omgeving zonder al te veel omgevingsgeluiden, want anders werd het geluid niet herkend, zoals mijn collega tot zijn verdriet menigmaal moest ervaren.

Slimme speakers

De laatste paar jaar gaan de ontwikkelingen op dit gebied razendsnel. Spraakassistenten als Siri en Google Assistant worden steeds populairder naarmate ze meer toepassingsmogelijkheden bieden. In Nederland kennen we ze vooral van de tablet en de smartphone, maar in Engelstalige landen zijn ze ook te vinden in zogenaamde 'slimme' speakers, die stormenderhand de markt veroveren.
Je kunt vanaf elke plek in je woonka-

mer handsfree opdrachten geven aan de speaker, waardoor het gebruik nog laagdrempeliger wordt. Het apparaat zet zijn microfoon(s) aan wanneer je een bepaalde triggertekst uitspreekt, bijvoorbeeld 'Hey Google'. De speaker heeft dezelfde functies als de spraakassistent op de smartphone: hij geeft het weerbericht, vertelt wat de hoogste berg van Afrika is, werkt je agenda bij, zet de wekker op half zeven en laat verzoeknummers horen. Daarnaast zijn er steeds meer slimme apparaten, zoals thermostaten, lampen en televisies, die 'luisteren' naar de slimme speaker.

Verkoopsucces

Het marktonderzoeksbureau comScore schat in dat er in 2020 meer gebruik gemaakt wordt van spraak dan van het toetsenbord. De belangrijkste aanjager van de ontwikkelingen op dit gebied is die van de commercie. Bedrijven zien de enorme mogelijkheden van spraaktechnologie om producten aan de man te brengen. Het is niet voor niets dat Amazon, als grootste internetverkoper, ook marktleider is op het gebied van spraakgestuurde apparaten.

Het paradepaardje is de 'slimme' speaker Echo, die werkt met de spraakassistent Alexa. De goedkoopste variant kost nog geen vijftig dollar, maar werd rond de feestdagen in december voor maar dertig dollar aangeboden. Amazon neemt het verlies dat het daarmee maakt voor lief, want het bedrijf levert zo niet alleen een handig apparaat aan miljoenen huishoudens, maar ook een directe toegang tot de eigen online store.

Alleen al vorig jaar zijn er meer dan dertig miljoen slimme speakers verkocht, waarvan ruim achttien miljoen in het laatste kwartaal. Meer dan de helft daarvan was van Amazon. De Google Home is in opmars en kwam met ruim dertig procent op een goede tweede plaats. Beide bedrijven doen er van alles aan om hun producten meerwaarde te geven. Zo geeft Amazon derden de mogelijkheid om toepassingen te maken die gebruikmaken van de Echo.

'Spraaktechnologie wordt ingezet bij kinderen met leesproblemen'

Spraaktechnologie heeft ook maatschappelijk nut

In de slipstream van de commerciële ontwikkelingen komen er ook steeds meer spraaktoepassingen die een meer maatschappelijk nut dienen. Voor dementiepatiënten is de spraakassistent bijvoorbeeld een geduldige gesprekspartner die er nooit genoeg van krijgt om steeds weer dezelfde antwoorden op dezelfde vragen te geven. Er bestaat spraaksoftware waarmee aandoeningen als Alzheimer en afasie in een vroeg stadium kunnen worden geconstateerd.

Spraaktechnologie wordt ook ingezet bij kinderen met leesproblemen. Hun gesproken woorden worden omgezet naar tekst die ze zelf weer moeten lezen. Omdat ze zich hun eigen woorden nog herinneren en omdat het woorden zijn uit hun eigen vocabulaire, zijn ze beter in staat om de tekst te lezen en te begrijpen waar het over gaat.

Blinden en slechtzienden, analfabeten en mensen met dyslexie kunnen dankzij text-to-speech-toepassingen informatie tot zich nemen waar ze voorheen geen toegang toe hadden.

Ondertussen zijn er al meer dan 26.000 van die zogenaamde skills. De Google Home, die later op de markt kwam, blijft daar met een kleine 2000 skills nog ruim bij achter.
In de slipstream van deze twee koplopers verschijnen er meer aanbieders op de markt. Apple is daarvan de meest bekende, maar hun HomePod blijft in slimheid achter bij de grote broers.

De techniek

De ontwikkeling van de techniek achter de spraakapparaten komt niet uit het niets. Al in de jaren vijftig ontwikkelde Bell Labs een systeem dat tien cijfers kon herkennen. Daarna ging het een tijd langzaam voorwaarts tot en met het dicteerprogramma dat mijn collega aan het begin van deze eeuw zoveel hoofdbrekens bezorgde. De enorme sprongen die er de afgelopen jaren zijn gemaakt, zijn te danken aan het gebruik van neurale netwerken, die losjes zijn gebaseerd op de manier waarop onze hersenen werken.
Wanneer een geluid ons oor binnenkomt, wordt dat opgedeeld in kleine stukjes die talloze keren door onze zenuwcellen worden verwerkt, waarbij er iedere keer wat informatie wordt verzameld. Uiteindelijk ontstaat er een totaalbeeld, waaruit we kunnen opmaken of iemand ons vriendelijk groet of dat er vlak achter ons een ballon is geknald. Ook in kunstmatige neurale netwerken worden gegevens in kleine stukjes steeds weer opnieuw verwerkt.
Het is een techniek die al tientallen jaren oud is, maar die enorm succesvol werd dankzij de toegenomen rekenkracht van computers, waardoor het nu mogelijk is om enorme hoeveelheden gesproken tekst op te slaan, te bewerken en te analyseren. Deze techniek, deep learning, onthoudt bijvoorbeeld welke woordcombinaties vaak voorkomen. Doordat een woord, in steeds andere contexten, miljoenen keren wordt verwerkt, 'leert' het systeem wat de betekenis van een woord is. Dit werkt vooral goed bij min of meer formele teksten.

Menselijk geluid

De techbedrijven doen hun uiterste best om hun spraakassistenten zo prettig en zo menselijk mogelijk te laten praten. Spraakgigant Nuance Communications laat straatnamen en rijrichting in navigatieapparatuur helder en gearticuleerd uitspreken, maar in dialogen spreekt hun spraakassistent vloeiender en meer dynamisch.
Meestal wordt er bij taal-naar-spraaksystemen gebruikgemaakt van een grote database met door een mens geproduceerde spraak. De software stelt teksten samen door kleine fragmentjes uit de database samen te voegen. Omdat alles zo goed mogelijk bij elkaar moet passen, zijn de klanken op neutrale toon ingesproken, waardoor er weinig dynamiek in de (toch uiteindelijk menselijke) stem zit.
Een ontwikkeling waar veel van wordt verwacht is spraak die volledig door computers wordt gegenereerd. Door de toegenomen computerkracht kunnen er veel meer eigenschappen van een klank worden meegegeven, waardoor er meer dynamiek en 'menselijkheid' in de mechanische stem kan worden opgenomen.
Hoe beter de spraaktechnologen erin slagen om hun apparaat als een mens te laten praten, hoe meer de echte mensen geneigd zullen zijn om met het apparaat te praten alsof ze het tegen een mens hebben. Daarmee plaatst de industrie zichzelf wel weer voor een nieuwe uitdaging.

Lastig probleem

Het grote voordeel van spraaktechnologie is natuurlijk het gemak en de snelheid waarmee je kunt communiceren met een computer of een door computertechnologie gestuurd apparaat. Je hoeft niet meer achter een toetsenbord te zitten, met een afstandsbediening te goochelen of op knoppen te drukken. De stem volstaat. Bovendien kan een mens gemiddeld 150 woorden per minuut spreken, terwijl hij er maar 40 kan typen. Een deel van dat snelheidsvoordeel gaat wel weer verloren omdat mensen over het algemeen meer woorden nodig hebben wanneer ze praten dan wanneer ze typen. Ook mensen die

'We zijn nog niet toe aan kantoortuinen waarin je mensen hoort roepen: Hey Google, print het jaarverslag even uit'

Zeg het als je het niet weet!

Toen ik Siri vroeg wat 'VOGIN' was, kreeg ik als antwoord de betekenis van *verhoging*.
Bij mijn volgende pogingen verstond Siri achtereenvolgens *for geen, vogel in, voor geen, Hoogeveen, van geen, volgend, Vogezen, vol geen, verhogen* en *volgen*. In mijn wanhoop legde ik mijn iPhone en mijn iPad naast elkaar en stelde mijn vraag tegelijkertijd aan beide Siri's. Mijn iPad verstond *voeg een* en mijn iPhone *voorin*. Ik deed een laatste check: de iPhone hoorde *verhoging* en de iPad *voor geen*.
De Google Assistant wist overigens evenmin raad met mijn zoekactie en ook nu verstond mijn iPhone iets anders dan mijn iPad. Gelukkig leverde een ingetypte zoekvraag in Google 451.000 resultaten op.

gewend zijn om in Google één of twee woorden in te tikken, hebben de neiging om een spraakassistent met een volzin te bevragen.
En naarmate de spraakassistenten zich 'menselijker' presenteren, zullen de vragen die ze krijgen ook steeds vaker in gewone mensentaal gesteld worden. Met 'Restaurants Amstelveen' heeft een spraakassistent weinig moeite. Anders is het wanneer ik zeg: 'Geef me een rijtje restaurants in, kom, wat was het ook alweer, ehm, uh, Amsterdam of Amstelveen? O ja, dat laatste.' En het wordt helemaal lastig wanneer ik die vraag stel in het West-Fries van mijn jeugd.
Behalve met slordig gesproken opdrachten hebben spraakassistenten ook moeite met niet-letterlijke taaluitingen, zoals ironie en sarcasme. Als ik zeg: 'Dat is lekker!' wanneer ik een vlieg midden in de soep zie drijven, zal een spraakassistent geneigd zijn te denken dat ik dol op insecten ben. Hoewel er op dit gebied stappen worden gezet, blijven spraaksystemen moeite houden met dergelijke meer subtiele taaluitingen.

'Spraaksystemen blijven moeite houden met meer subtiele taaluitingen'

Gevaarlijke dolfijnen

Gaat zoeken via spraak ons alleen maar voordeel opleveren? Nee, natuurlijk niet, al is het maar omdat elke nieuwe internettoepassing ook weer nieuwe veiligheidsrisico's met zich meebrengt. De woorden die je richt tot je slimme speaker blijven daar niet hangen, maar worden doorgestuurd naar de servers van de leverancier, waar de tekst verder wordt verwerkt.
De grote firma's roepen om het hardst dat de tekst niet wordt opgeslagen, maar wie garandeert dat de gegevens zo goed beveiligd worden dat er niemand bij kan? En hoe zit het met de microfoons in de slimme speakers? Nemen die echt alleen geluid op als je daar zelf voor kiest of kunnen ze door kwaadwillenden zo worden gemanipuleerd dat ze de hele dag meeluisteren?
Misschien biedt de Mycroft Mark II binnenkort een oplossing. De ontwikkelaars van deze slimme speaker maken gebruik van open source software en claimen dat privacy en data-onafhankelijkheid gewaarborgd zijn. De toekomst zal het uitwijzen, maar wel is duidelijk dat spraakgestuurde apparaten kwetsbaar zijn. Zo ontdekten onderzoekers dat het mogelijk is om ultrasone, voor mensen onhoorbare signalen, naar spraakapparatuur te sturen. Via deze 'Dolphin attacks' slaagden ze er bijvoorbeeld in om de speaker foto's te laten maken.
De FBI waarschuwt nadrukkelijk voor 'smart toys' die zijn voorzien van spraakherkenning. Kinderen zijn makkelijke slachtoffers. In Duitsland is dit jaar een pop uit de handel genomen die een eenvoudig te hacken microfoontje bevatte en de Noorse consumentenbond toonde zelfs aan dat het mogelijk is om op die manier via de pop met kinderen te praten.
Maar er zijn ook risico's op een ander vlak. Wanneer we nu iets opzoeken via Google, krijgen we een scherm met opties waaruit we zelf één of meerdere keuzes kunnen maken. Een slimme speaker zal in zo'n geval vaak maar één antwoord geven en we zullen geneigd zijn om daarop af te gaan. Marketeers hebben er een dagtaak aan om ervoor te zorgen dat juist hun product op die eerste plaats komt.

Nederland is (nog) te klein

De slimme speakerhype ging tot nu toe grotendeels aan ons land voorbij. Dat komt omdat we tot een klein taalgebied behoren en de techbedrijven zich voorlopig vooral richten op de Engelstalige markt. In Nederland is alleen nog de 'International Version' van de Echo verkrijgbaar. Deze heeft beduidend minder functies dan de Amerikaanse variant. Google komt binnenkort met een spraakassistent die Nederlands spreekt, en naar verluidt werkt het bedrijf ook aan een Nederlandstalige Google Home, maar Google hult zich hierover tot nu toe in stilzwijgen.
Ondertussen staan bedrijven te popelen om gebruik te maken van de slimme speakers. Zo onderzoeken verschillende supermarkten hoe ze hun klanten via stemopdrachten boodschappen kunnen laten bestellen en heeft nieuwssite NU.nl al software gebouwd die via de slimme speaker nieuwsberichten kan voorlezen.
Het is een kwestie van tijd dat voice search, in de vorm van slimme speakers, de Nederlandse markt zal veroveren. Net zoals het een kwestie van tijd is dat ze, op een iets langere termijn, ons leven gaan veranderen. Zeker in het begin zal dat vooral in de privésfeer zijn.
We zijn nog niet toe aan kantoortuinen waarin je mensen hoort roepen 'Hey Google, print het jaarverslag even uit'. En vooralsnog zullen de speakers vooral de opdrachten uitvoeren die we voorheen via het toetsenbord gaven. In die zin kunnen we misschien beter spreken van dienstbare speakers. Het tijdperk van de écht slimme speaker ligt nog ver voor ons. Maar laten we niet vergeten dat we die geluiden de laatste jaren wel vaker hebben gehoord over ondertussen verwezenlijkte utopieën. <