

Hoor wie het zegt

Handvatten voor het verantwoorde gebruik van spraaktechnologie



Auteurs

Hamer, J., S. Doesborgh en L. Kool

Redactie

Rathenau Instituut

Illustraties en foto's

L. Marienus

Foto omslag

ANP

Bij voorkeur citeren als:

Rathenau Instituut (2020). *Hoor wie het zegt – Handvatten voor het verantwoorde gebruik van spraaktechnologie*. Den Haag (auteurs: Hamer, J., S. Doesborgh en L. Kool)

Voorwoord

De afgelopen jaren is spraaktechnologie gemeengoed geworden. Vele chauffeurs geven mondelinge instructies aan hun auto's en sommige mensen staan 's ochtends zelfs op met de stem van een digitale spraakassistent, die de gebruiker goedemorgen wenst en de weersvoorspelling meegeeft. We gaan steeds vaker met computers een gesprek aan – en dat heeft consequenties.

Digitale technologie heeft de afgelopen decennia een ingrijpende impact gehad op onze cultuur en sociale normen. Denk aan de etiquette van video calls en de debatten op sociale media. Spraaktechnologie zal onze sociale relaties en normen nog verder beïnvloeden. Niets is immers menselijker dan onze spraak. In onze gesprekken drukken we onszelf uit en ontwikkelen we omgangsvormen. Het is daarom van belang om spraaktechnologie in juiste banen te leiden.

Daarnaast komt spraaktechnologie dichtbij ons: we installeren spraaksystemen in onze woonkamers en op onze kantoren. In de verkeerde handen is een spraakcomputer een surveillance-instrument waarmee onze geheimen ontsloten kunnen worden. Je kan zelfs stemmen klonen en iemand woorden in de mond leggen. Bovendien staat onze zelfbeschikking op het spel. Spraaktechnologie functioneert steeds meer als een gids die de gebruiker door de digitale wereld leidt. Maar deze gids wordt gemaakt door bedrijven die hun eigen belangen nastreven, en die sluiten niet per se aan bij de belangen en wensen van burgers.

Het Rathenau Instituut heeft deze studie daarom gewijd aan spraaktechnologie. We laten op basis van deskresearch en interviews zien hoe spraaktechnologie werkt, waarvoor zij gebruikt wordt en welke ethische vragen zij oproept. We zoeken uit hoe de overheid, bedrijven en burgers kunnen bijdragen aan spraaktechnologie die onze samenleving en sociale relaties verrijkt, en niet verschaalt.

De studie roept op tot het ontwikkelen van ethische spraaktechnologie die, onder meer, inclusief is, ons privéleven respecteert en aangeboden wordt op een gezonde markt. De studie roept ook op tot maatschappelijke dialoog en politiek debat. De opkomst van spraaktechnologie brengt vragen met zich mee die we samen moeten beantwoorden. Willen we bijvoorbeeld gedisciplineerd worden door een spraakassistent? Voorheen zou deze vraag fantasieerijk hebben geklonken, maar vandaag is hij reëel. Computers zijn gaan praten: tijd voor een goed gesprek.

Dr. ir. Melanie Peters
Directeur Rathenau Instituut

Samenvatting

Inleiding

Computers kunnen steeds beter menselijke spraak herkennen, interpreteren en produceren. Dankzij de verbeterde spraaktechnologie is het mogelijk om met computers in gesprek te gaan, en kunnen gebruikers met hun stem de digitale wereld aansturen. Spraaktechnologie wordt al veel gebruikt in de auto en in huis, en bedrijven en organisaties experimenteren ermee in tal van andere domeinen, waaronder de zorg en de veiligheidssector. De groeiende toepassing van spraaktechnologie heeft belangrijke gevolgen voor de samenleving. Onze spraak vormt een essentieel onderdeel van wie we zijn als mens, en van onze sociale relaties. Ook bevatten onze gesprekken zeer gevoelige informatie – over onze identiteit, over de gesprekken die we voeren, en zelfs over onze gezondheid en ons humeur. Onze spraak verdient dus bescherming. Deze studie onderzoekt hoe de samenleving deze bescherming vorm kan geven.

De studie is gebaseerd op deskresearch en interviews. De deskresearch bestond uit het bestuderen van een combinatie van academische literatuur en grijze literatuur, gezien de relatief recente doorbraak van spraaktechnologie. De interviews waren verkennend van aard, en bedoeld om een beter beeld te krijgen van de technische mogelijkheden van spraaktechnologie.

Spraaktechnologie wordt steeds beter

In deze studie analyseren we allereerst de technische stand van zaken: hoe werkt spraaktechnologie, en hoe goed is het? Spraaktechnologie bestaat uit drie centrale processen: het herkennen van spraak, het interpreteren van spraak, en het spreken zelf, dat spraaksynthese wordt genoemd (zie figuur 1). Op alle drie de terreinen is vooruitgang geboekt, met name dankzij rijkere en grotere datasets, geavanceerde *machine learning*-technologie en de snellere rekenkracht van computers. Maar ondanks de kwaliteitsverbetering is het beeld gemengd.

Figuur 1 Elementen van spraaktechnologie



Spraakherkenning werkt al behoorlijk goed. Spraakcomputers behalen, in ideale omstandigheden, een foutmarge van rond de 5%. Maar de omstandigheden maken veel uit: de foutmarge neemt sterk toe op een rumoerige plek, als er technische woorden worden gebruikt of als het systeem luistert naar stemmen van groepen die minder sterk in de trainingsdata vertegenwoordigd zijn, zoals die van kinderen. Toch is de spraakherkenning voldoende nauwkeurig om tal van nuttige diensten te verlenen, bijvoorbeeld als het gaat om het op afstand bedienen van muziek of het transcriberen van een interview. Maar er zijn genoeg toepassingen te bedenken, zoals in de zorg of zware industrie, waarbij een dergelijke foutmarge niet acceptabel is.

Op het gebied van **spraakinterpretatie** is de voortgang minder eenduidig. Bij het uitvoeren van taken is hulp van de omgeving en de gebruiker nodig: hij of zij moet simpele commando's geven en vragen op de juiste manier formuleren en beantwoorden. Hoewel de belofte was dat computers onze taal gingen leren, moet een mens zich nog steeds aanpassen aan de spraaktechnologie, wil hij goed geïnterpreteerd worden.

De **spraaksynthese** is juist veel beter geworden. Kortgezegd kunnen spraaksystemen zich al duidelijk verstaanbaar maken. Ontwikkelaars hebben de lat hoger gelegd: spraaksynthese moet zo goed zijn, dat mensen niet meer horen dat ze met een computer spreken. Bij verreweg de meeste toepassingen is dit nog niet het geval, maar de ontwikkelingen gaan hard. Sommige spraaksystemen, zoals Google Duplex, komen heel dichtbij het produceren van menselijke spraak, inclusief 'uhhs' en 'mmhmmh's'.

Spraaktechnologie is onze gids in de digitale wereld

Deze studie heeft ook de toepassing van spraaktechnologie in kaart gebracht. Met name in de auto en thuis wordt spraaktechnologie al breed toegepast. Daarnaast experimenteren technologieaanbieders en bedrijven volop met spraak. De toepassingen kan je onderverdelen in twee groepen: toepassingen die **apparaten aansturen** en toepassingen die **diensten ondersteunen of overnemen**. In de eerste categorie kennen we spraaktechnologie in de auto (handsfree bellen) en in huis (de spraakassistenten, zoals Google Assistant of Alexa van Amazon). Maar ook in de industrie kunnen machines via spraak worden aangestuurd. In de tweede categorie zien we onder meer spraakassistenten die reizen voor ons boeken, ons assisteren op kantoor, en onze identiteit checken, bijvoorbeeld als we bankzaken willen regelen.

Figuur 2 Ethische aspecten van spraaktechnologie



Effecten op sociale relaties en normen

Deze brede waaier aan toepassingen brengt verschillende maatschappelijke en ethische vraagstukken met zich mee (zie figuur 2). Ten eerste mengt spraaktechnologie zich in het sociale leven van mensen. Dit roept vragen op over

de gewenste relatie tussen mensen en computers: willen we, en moeten we altijd weten, dat we tegen een computer spreken in plaats van een mens? Horen we wel wie, of wat, iets zegt? Is het problematisch als gebruikers hun spraakassistent beschouwen als hun beste vriend? En hoe zorgen we dat spraaktechnologie bestaande sociale normen, bijvoorbeeld ten aanzien van gelijke behandeling en disciplineren, respecteert? We moeten ervoor waken dat spraaktechnologie onze waardigheid als mens onder druk zet.

De stem als nieuwe databron

Bovendien verzamelen al deze toepassingen data, door middel van gespreksverslagen én audio-opnames. Onze studie laat zien dat de stem zo fungeert als nieuwe databron. De data worden door ontwikkelaars gebruikt om spraaksystemen te personaliseren, en vormen de basis van analyses op het gebied van emotieherkenning en het diagnosticeren van ziekten. Die analyses zijn vaak niet wetenschappelijk bewezen, maar diverse bedrijven verwachten veel van de toekomstige mogelijkheden van audio-opnamen. Spraakdata bevatten zeer gevoelige data: juist in gesprekken thuis, in de auto en op het werk geven mensen zichzelf bloot. Dat vraagt om extra aandacht van ontwikkelaars en toezichhouders, zodat ons privé- en familieleven gerespecteerd blijft.

Onze autonomie staat op het spel

De inzet van spraaktechnologie heeft ook invloed op onze autonomie. De technologie helpt in tal van domeinen om taken uit te voeren, beslissingen te nemen en gebruikers een aangename ervaring te geven. Dit biedt kansen, maar brengt ook zorgen met zich mee. Leidt het gebruik van spraaktechnologie tot het verlies van vaardigheden, en vinden er oneigenlijke beïnvloeding en sturing plaats? Denk bijvoorbeeld aan *deep fake* video's, waarin iemands uiterlijk en stem worden nagemaakt, ('gekloond'), en die mensen om de tuin kunnen leiden en het publieke debat ondermijnen. Daarnaast geeft spraaktechnologie in vergelijking met schermen minder mogelijkheden voor nuance en doorvragen. Wie stuurt en bepaalt welk antwoord de spraakassistent geeft? Ten slotte kan een empathische en handige spraakassistent zo nuttig zijn, dat mensen er te veel gebruik van maken en verslaafd raken.

Het belang van veilig en gezond gebruik

Spraaktechnologie kan ook de veiligheid van mensen onder druk zetten. Spraakdata kan gestolen en misbruikt worden, bijvoorbeeld om identiteitsfraude mee te plegen. En ondanks de verbeteringen is spraaktechnologie niet foutloos en kunnen er ongelukken gebeuren. Voordat spraaktechnologie in kritieke toepassingen in de zorg, defensie of de maakindustrie wordt ingezet, zal de betrouwbaarheid van de technologie buiten kijf moeten staan, en geïnvesteerd moeten worden in technieken die misbruik juist tegengaan.

Groeiende marktmacht van techgiganten

De studie laat ten slotte zien dat de macht van diverse grote technologiebedrijven via spraaktechnologie nog verder aan het groeien is. De doelstelling van verschillende technologiegiganten zoals Google en Amazon is om een breed platform van spraaktoepassingen te creëren, en die te koppelen aan een spraakassistent, zoals Alexa en Google Assistant, die een veelheid aan taken kan uitvoeren. Daarmee nemen deze assistenten de rol aan van een gids die ons door de digitale wereld leidt, en ons zoveel mogelijk binnen de omgeving van een bepaald platform houdt. Om dit te bereiken kopen technologiegiganten daarbij start-ups op, en doen significante investeringen. Hoewel er ook andere partijen actief zijn op de markt voor spraaktechnologie, zoals het platform Houndify, en bedrijven soms zelfstandig spraakassistenten ontwikkelen, is het de vraag hoe deze partijen zich staande houden ten opzichte van de groeiende machtspositie van de techgiganten.

Aanbevelingen voor de overheid en het bedrijfsleven

Onze stemmen en gesprekken vormen een essentieel onderdeel van wie we zijn als mens, en van de relaties die we met anderen aangaan. Met spraaktechnologie krijgen we er op allerlei momenten – thuis, in de auto, op het werk en bij het winkelen – een gesprekspartner bij, en dat zal onze spraak en onze relaties – met elkaar én met computers – niet onberoerd laten. Bovendien creëert spraaktechnologie een nieuwe bron van data, die zeer gevoelige gegevens bevatten. Onze spraak staat op het spel.

Spraaktechnologie voegt een nieuwe dimensie toe aan de brede opdracht om digitale technologie in goede banen te leiden, en laat zien dat de overheid en het bedrijfsleven wederom aan zet zijn. Spraaktechnologie beïnvloedt namelijk niet alleen de individuele omgang met computers, maar ook de omgangsvormen die we samen ontwikkelen. Zij verandert niet alleen de manier waarop individuen kennis vergaren, maar ook de kennisbasis van het publieke debat. En zij heeft niet alleen impact op de relatie tussen klanten en bedrijven, maar ook op de platformeconomie als geheel.

Het Rathenau Instituut komt daarom met zes aanbevelingen aan de overheid en het bedrijfsleven om menselijke spraak te beschermen en de toepassingen van spraaktechnologie in juiste banen te leiden:

Aanbevelingen	Overheid	Bedrijfsleven
1. Zorg voor effectieve privacybescherming	Voer vergunningsplicht voor stemanalyse in. Monitor gebruik spraaktechnologie door opsporingsdiensten.	Geef privacybeginselen krachtig vorm
2. Bevorder inclusieve spraaktechnologie	Investeer in Nederlandse spraakdatabase. Spreek bedrijven aan op hun verantwoordelijkheden.	Waak voor stereotypering in stemgebruik en bevorder herkenning divers taalgebruik.
3. Creëer een eerlijke markt	Scherp het mededingingsrecht aan. Bied mogelijkheden voor alternatieve aanbieders.	Zet de rechten van consumenten centraal.
	Overheid en bedrijfsleven	
4. Bescherm menselijke waardigheid	Start ethische dialoog over spraaktechnologie en kom tot afspraken.	
5. Zorg dat spraaktechnologie betrouwbaar is	Pak desinformatie aan. Verlaag de foutmarges van spraaksystemen.	
6. Investeer in technologisch burgerschap	Leid burgers op om verantwoord om te gaan met spraaktechnologie, en stimuleer onderzoek.	

1. Zorg voor effectieve privacybescherming

Spraaktechnologie maakt het mogelijk om gevoelige stemdata van mensen te verzamelen en te gebruiken om mensen te beïnvloeden. Het gaat onder meer om biometrische gegevens en gegevens over gezondheid. De verwerking van stemdata levert zo risico's op voor de mens en diens fundamentele rechten. Het is zaak de bestaande privacyregels krachtiger te handhaven. Daarom roept het Rathenau Instituut de overheid op om een vergunningsplicht voor biometrische stemanalyse in te voeren en strategieën te ontwikkelen om emotieherkenning en gezondheidsanalyse te reguleren. Ook is het van belang om de inzet van spraakanalyse door opsporingsdiensten te monitoren: is het wenselijk dat de politie stemdata van sociale media schraapt? Ten slotte is het aan het bedrijfsleven om in hun productontwikkeling en dienstverlening de privacyregels niet minimaal te volgen, maar krachtig vorm te geven – bijvoorbeeld door te investeren in technieken die het gebruik van data minimaliseren.

2. Bevorder inclusieve spraaktechnologie

Spraaktechnologie biedt kansen om informatie makkelijker toegankelijk te maken. Maar spraaksystemen kunnen ook groepen gebruikers uitsluiten, vooroordelen bevestigen, en discriminatie in de hand werken. Het is van groot belang om ervoor te zorgen dat iedereen spraaktechnologie kan benutten. De overheid kan daartoe investeren in een Nederlandse spraakdatabase, waar tal van partijen hun

spraaktechnologie op kunnen baseren. Ook het bedrijfsleven heeft hier verantwoordelijkheden. Het Rathenau Instituut roept het bedrijfsleven in het bijzonder op om stereotypering tegen te gaan, bijvoorbeeld door een diverse reeks spraakassistenten aan te bieden.

3. Creëer een eerlijke markt

Er bestaan zorgen in de dataeconomie over de dominantie van enkele grote technologiebedrijven. Spraaktechnologie biedt deze bedrijven de mogelijkheid om deze machtspositie nog verder uit te breiden. Om de markt voor alle partijen toegankelijk en eerlijk te maken, kan de overheid het mededingingsrecht aanscherpen – daartoe worden op het Europese niveau stappen gezet. Ook is het zaak om mogelijkheden te bieden voor alternatieve aanbieders, en niet alleen samen te werken met de techgiganten. Het bedrijfsleven wordt aanbevolen om de rechten van consumenten, zoals het recht om informatie op te vragen, slagkrachtig en ruimhartig vorm te geven.

4. Bescherm menselijke waardigheid

Het Rathenau Instituut roept de overheid en het bedrijfsleven op om een ethische dialoog over spraaktechnologie te starten. Daarbij zou bijzondere aandacht uit moeten gaan naar het beschermen van menselijke waardigheid: het borgen van het recht op menselijk contact en het voorkomen van situaties waarin gebruikers computers en mensen met elkaar verwarren. Daarover zouden de overheid en het bedrijfsleven met elkaar afspraken moeten maken.

5. Zorg dat spraaktechnologie betrouwbaar is

Spraaktechnologie kan de samenleving veel bieden, maar moet wel betrouwbaar zijn. Het is zowel aan de overheid als aan het bedrijfsleven om hiertoe de volgende stappen te zetten: treed slagvaardig op tegen desinformatie en *voice cloning*, verlaag de foutmarge van spraaktechnologie, investeer in technologie die misbruik tegengaat en ontwikkel veiligheidsstandaarden.

6. Investeer in technologisch burgerschap

Een verantwoord en effectief gebruik van spraaktechnologie vereist ook kennis en vaardigheden, bijvoorbeeld ten aanzien van het zoeken naar kennis en het instellen van routines, en de informatie die de apparaten verzamelen. Het is daarom nodig burgers te ondersteunen bij de omgang met spraaktechnologie. Dit vereist investeringen in opleidingen en trainingen op het gebied van mediawijsheid. Daarnaast zijn van de overheid, kennisinstellingen en het bedrijfsleven investeringen in onderzoek nodig om de invloed van spraaktechnologie op onze fysieke en mentale gezondheid te analyseren. Ten slotte is er ook een belangrijke rol weggelegd voor burgers. Zij kunnen hun stem laten horen en spraaktechnologie

op de agenda van het publieke debat zetten. Onze spraak is een kwetsbaar en betekenisvol goed – en een discussie meer dan waard.

Inhoud

Voorwoord.....	3
Samenvatting	4
Inleiding	14
1.1 Computers die spreken.....	14
1.2 Afbakening.....	16
1.3 Onderzoeksvragen	16
1.4 Onderzoeksmethode	17
2 Spraaaktechnologie: een technische analyse	18
2.1 Spraaaktechnologie in een notendop	18
2.2 Spraaakherkenning	21
2.2.1 Hoe werkt spraaakherkenning?	21
2.2.2 Hoe goed werkt spraaakherkenning?	23
2.2.3 Hoe werkt spraaakinterpretatie?	24
2.2.4 Hoe goed werkt spraaakinterpretatie?.....	27
2.3 Spraaaksynthese	28
2.3.1 Hoe werkt spraaaksynthese?	28
2.3.2 Hoe goed werkt spraaaksynthese?.....	29
2.4 De toekomst van spraaaktechnologie	31
2.5 Conclusie.....	32
3 Waar wordt spraaaktechnologie voor gebruikt?	34
3.1 De toepassingen die spraaaktechnologie kan uitvoeren	34
3.1.1 Apparaten aansturen.....	35
3.1.2 Dienstverlening ondersteunen, of overnemen	38
3.2 Data-analyse door bedrijven.....	43
3.2.1 Tekstanalyse.....	43
3.2.2 Stemanalyse	44
3.3 De markt voor spraaaktechnologie	45
3.4 Conclusie.....	50
4 Maatschappelijke en ethische aspecten van spraaaktechnologie	52
4.1 Het belang van spraaak	52
4.2 Sociale relaties en normen	54
4.3 Respect voor privacy	58
4.4 Autonomie	61

4.5	Veilig en gezond gebruik	64
4.6	Macht van technologiebedrijven	66
4.7	Conclusie	67
5	Conclusie: tijd voor een goed gesprek.....	68
5.1	De stand van zaken	68
5.1.1	Spraaktechnologie wordt volwassen.....	68
5.1.2	Spraaktechnologie wordt steeds meer gebruikt	69
5.1.3	Maatschappelijke en ethische aspecten van spraaktechnologie	70
5.2	Spraaktechnologie vraagt om maatschappelijke en politieke actie.....	72
1.	Zorg voor effectieve privacybescherming.....	73
2.	Bevorder inclusieve spraaktechnologie	75
3.	Creëer een eerlijke markt.....	76
4.	Bescherm menselijke waardigheid	77
5.	Zorg dat spraaktechnologie betrouwbaar is	77
6.	Investeer in technologisch burgerschap.....	78
5.3	Slotwoord.....	79
	Literatuur	80

Inleiding

1.1 Computers die spreken

Soms wordt sciencefiction werkelijkheid. Schrijvers en filmmakers zoals Isaac Asimov – I, Robot – en Stanley Kubrick – 2001: A Space Odyssey – fantaseerden decennia geleden al over computers die kunnen spreken als mensen. Die fantasie is de afgelopen jaren steeds meer realiteit geworden.

Door de opkomst van kunstmatige intelligentie en de toegenomen rekenkracht van processors kunnen computers steeds beter spraak herkennen, interpreteren en spreken. De menselijke taal is niet langer uitsluitend het domein van mensen – in steeds meer gesprekken doen spraaksystemen met ons mee. Zoals Google Duplex, die restaurants voor je opbelt en namens jou een reservering maakt. Of Microsoft Cortana, waarmee je je agenda kan beheren en e-mails kan versturen. Of Athena, een spraakassistent voor de industrie, waarmee machines met de stem zijn aan te sturen.

De markt voor spraaktechnologie groeit momenteel snel. Zo heeft 6% van de Nederlandse huishoudens in 2018 een luidspreker aangeschaft die je met spraak kunt bedienen, en groeide dit percentage in 2019 tot 19% (Multiscope, 2020). En in Amerika en China gaan de ontwikkelingen nog harder (Kimmich, 2019). Volgens sommige analyses lijkt de opkomst van slimme luidsprekers daar zelfs sneller te gaan dan destijds de opkomst van mobiele telefoons – een apparaat dat je tegenwoordig ook steeds vaker met je stem kunt besturen (Kinsella & Mutchler, 2018).

De interesse in spraaktechnologie groeit, omdat je met een spraaksysteem snel en intuïtief kan interacteren. Snel, omdat mensen gemiddeld 150 woorden per minuut kunnen zeggen, maar slechts veertig woorden per minuut kunnen typen. Intuïtief, omdat wij voor het eerst niet de taal van computer zouden hoeven te leren, maar computers zich onze taal eigen maken.

De opkomst van spraaktechnologie zal belangrijke impact hebben op de samenleving. Onze relatie tot de digitale wereld verandert als we tegen computers – en allerlei digitale apparaten – kunnen praten. Waar we voorheen via zoekmachines op een scherm zochten en doorklikten, klinkt er nu één antwoord terug. Waar we voorheen de klantenservice belden om geholpen te worden door

een mens, kun je nu een digitale spraakassistent aan de lijn krijgen, die steeds meer als een mens klinkt.

Ook onze relatie tot elkaar verandert met spraaktechnologie. In een gezin hoeft niet iedereen op zijn eigen scherm te kijken, maar kun je samen een gesprek voeren met de spraakassistent. En misschien beïnvloedt de manier waarop we met computers praten ook de manier waarop we met elkaar spreken.

Ten slotte verandert onze relatie tot bedrijven en instellingen. Bijvoorbeeld doordat we niet langer met een menselijke medewerker in contact komen, maar met een digitale spraakassistent van een bedrijf spreken. Of doordat in transacties met bedrijven onze stem als middel gebruikt wordt om onze identiteit te controleren. Bedrijven gaan zowel zien *wat* we tegen computers zeggen als *hoe* we dat zeggen: onze stem, gesprekken en emoties vormen een nieuwe, interessante bron van data die bedrijven kunnen analyseren. Denk bijvoorbeeld aan het detecteren van ziektes en het identificeren van personen.

Onze spraak, en onze gesprekken, zijn kostbaar. Met onze taal drukken we uit wie we zijn, en gaan we talrijke waardevolle relaties aan. Niets is meer intiem, en meer eigen, dan onze stem. Daarom is het belangrijk om, gezien de opkomst van spraaktechnologie, nu te onderzoeken welke maatschappelijke en ethische vragen met spraakcomputers verbonden zijn. Wat betekent het gebruik van spraaktechnologie voor onze privacy? Wat gebeurt er als we van kinds af aan leren commando's te geven aan digitale apparaten? Kunnen we spraakcomputers en mensen nog wel uit elkaar houden? Wat zijn de gevolgen als spraakcomputers ons gidsen door de digitale wereld? En hoe zal spraaktechnologie de platformeconomie, en de concurrentiestrijd tussen technologiegiganten als Google en Amazon, beïnvloeden?

Deze verkenning zal op deze en andere vragen ingaan. We zullen een overzicht geven van de technologische ontwikkelingen, inventariseren waar spraaktechnologie gebruikt wordt en de maatschappelijke en ethische aspecten in kaart brengen. De verkenning zal laten zien waarom een maatschappelijk en politiek gesprek over spraaktechnologie nodig is, en waar dat gesprek over moet gaan. De verkenning doet aanbevelingen voor de overheid en voor het bedrijfsleven, en identificeert centrale thema's voor het publieke debat.

Isaac Asimov voorzag dat intelligente robots regels en grenzen nodig hadden. Hij bedacht daarom drie regels, waarvan de eerste de belangrijkste was: 'Een robot mag een mens geen letsel toebrengen of door niet te handelen toestaan dat een mens letsel oploopt' (Asimov, 2008). Nu een nieuw stukje sciencefiction werkelijkheid wordt, zal ook onze samenleving moeten nadenken over de publieke

waarden en normen die de opkomst van spraaktechnologie in goede banen kunnen leiden.

1.2 Afbakening

Deze verkenning gaat over spraaktechnologie. Hieronder verstaan we technologie die ervoor zorgt dat computers menselijke spraak kunnen herkennen, kunnen interpreteren en ook zelf kunnen spreken – samengevat als spraakherkenning, spraakinterpretatie en spraaksynthese (zie hoofdstuk twee).

Een computersysteem dat één of meerdere van deze drie taken kan uitvoeren, noemen we een **spraakstelsel**. Er zijn op de markt verschillende soorten spraaksystemen beschikbaar. De belangrijkste is de **spraakassistent**, een spraakstelsel dat doorgaans een breed aanbod van taken uit kan voeren. Bekende voorbeelden zijn Alexa van Amazon en de spraakassistent van Google. Deze assistenten kunnen op allerlei digitale apparaten, zoals een mobiele telefoon, een desktop pc, of een slimme luidspreker – een **smart speaker** – geïnstalleerd worden. De verkenning kijkt daarnaast ook naar andere spraaksystemen, zoals transcriptiesoftware en navigatiesystemen.

Spraakassistenten worden ook wel **cognitieve** of **virtuele assistenten** genoemd. Deze digitale systemen kunnen ook taken uitvoeren, en zijn doorgaans in staat om tekst te interpreteren. Ze hoeven niet gebaseerd te zijn op spraaktechnologie. In deze verkenning richten wij ons op systemen uitgerust met spraaktechnologie. We zullen daarom de term ‘spraakassistent’ gebruiken.

1.3 Onderzoeksvragen

Spraaktechnologie is met een snelle opmars bezig. De centrale onderzoeksvraag van deze verkenning luidt daarom:

Hoe kan de opkomst van spraaktechnologie in goede banen worden geleid?

Om deze vraag in kaart te brengen, voert deze studie drie analyses uit. Allereerst analyseren we de **technische mogelijkheden** van spraaktechnologie. Hoe heeft spraaktechnologie zich de afgelopen jaren ontwikkeld? Wat is er nu mogelijk, wat niet, en hoe verwacht men dat de toekomst van spraaktechnologie eruit ziet?

Vervolgens analyseren we **welke toepassingen** spraaktechnologie heeft, en welke in opkomst zijn. Waar wordt spraaktechnologie voor gebruikt? En door wie? Wat

gebeurt er met de data die spraaktechnologie verzamelt? Wat valt er uit een stem af te leiden? En hoe ziet de markt voor spraaktechnologie eruit?

Ten slotte analyseren we de **maatschappelijke** en **ethische aspecten** die verbonden zijn met de opkomst van spraaktechnologie. Hoe verandert spraaktechnologie de interactie tussen mensen en apparaten, en tussen mensen onderling? Welke vragen roept dat op? Welke publieke waarden zet spraaktechnologie onder druk, en welke publieke waarden kunnen door spraaktechnologie worden bevorderd?

Op basis van deze drie analyses, presenteren we in de conclusie een aantal aanbevelingen en identificeren we thema's voor het publieke debat.

1.4 Onderzoeksmethode

Deze verkenning is gebaseerd op deskresearch en interviews.

De deskresearch bestond uit een combinatie van academische literatuur en grijze literatuur, gezien de relatief recente doorbraak van spraaktechnologie. Grijze literatuur verwijst hier naar niet-academische bronnen zoals overheidsdocumentatie en -rapporten, journalistieke artikelen en publicaties uit de private sector. De interviews waren verkennend van aard, en bedoeld om een beter beeld te krijgen van de technische mogelijkheden van spraaktechnologie. Ze werden gehouden met experts op het gebied van spraaktechnologie: Roeland Ordelman van TU Twente, Olya Kudina van de TU Delft, Jeroen Vonk van Novum, Piet van Dosselaar van het Centraal Bureau voor Statistiek, Voice consultant Maarten Lens-FitzGerald, Vanessa Hendriks van Google Nederland, Jeff Adams van het bedrijf Cobalt Speech en Clive Summerfield van het bedrijf Auraya Systems.

2 Spraaktechnologie: een technische analyse

Als we technologiebedrijven moeten geloven, zijn de beloftes van spraaktechnologie groot. Technologiebedrijven zien in spraak een centraal middel om vanuit je luie stoel de digitale wereld, met al zijn mogelijkheden, te besturen. Zie bijvoorbeeld het *Voice Report* van Microsoft (Olson & Kemery, 2019). Met je stem schrijven, gegevens opvragen, muziek aanzetten, deuren openen, auto's rijden, bouwmachines aansturen – spraak is volgens hen de *interface* van de toekomst.

Of deze droom uitkomt, hangt onder meer af van hoe goed de technologie werkt.¹ Dit hoofdstuk brengt de technologische ontwikkeling van spraaktechnologie in kaart. We leggen uit wat spraaktechnologie precies is en onderzoeken hoe spraaktechnologie verschillende toepassingen mogelijk maakt. We onderzoeken ook de toekomstige beloftes van spraaktechnologie – welke toepassingen liggen nu buiten bereik, maar zouden in de nabije toekomst gerealiseerd kunnen worden?

Dit hoofdstuk is met name gebaseerd op het werk van computerwetenschappers die onderzoek doen naar spraaktechnologie.

2.1 Spraaktechnologie in een notendop

Spraaktechnologie laat de gebruiker toepassingen aansturen door zijn of haar stem te gebruiken. Als je bijvoorbeeld benieuwd bent naar het weer van vandaag, type je niets in, maar vraag je direct hardop aan een apparaat, zoals een smartphone of een smart speaker, wat het weer wordt. Het is gelijk helder wat het voordeel is van spraaktechnologie: praten is voor de meeste mensen sneller en makkelijker dan typen. Dit laat wetenschappelijk onderzoek naar de bediening van apparaten ook zien (Nass & Brave, 2005).

In deze studie onderscheiden we drie onderdelen van spraaktechnologie. Ten eerste moet de spraak **herkend** worden. Het systeem moet de opgevangen geluiden omzetten in tekst – daarom noemt men deze technologie ook wel *speech*

¹ Hoe een technologie uiteindelijk gebruikt wordt, hangt af van vele factoren, onder meer de manier waarop gebruikers zich de technologie zullen toe-eigenen, en de manier waarop partijen met diverse maatschappelijke en ethische kwesties omgaan. De volgende hoofdstukken zullen nader ingaan op deze factoren.

to text (Liu, e.a., 2019). Van alles kan hierbij misgaan: verschillen tussen woorden, zoals ‘kerk’ en ‘perk’, zijn soms klein. En door ruis kunnen woorden verloren gaan.

Vervolgens moet het systeem de opgevangen tekst **interpreteren** (Mittal, 2019). Als iemand de vraag stelt ‘hoe oud is de domtoren van Utrecht?’, is het cruciaal dat de software deze woorden op de juiste manier, en in het juiste verband, uitlegt. De gebruiker wil bijvoorbeeld niet weten hoe oud de stad Utrecht is – en dat moet het systeem begrijpen.

Vervolgens moet het systeem het gevraagde uitvoeren. Dit kan om van alles gaan: van het opzoeken en afspelen van een muziknummer tot het bellen van mensen en het hoger zetten van de verwarming. Soms sluit het apparaat de interactie af met een bevestiging zoals: ‘Prima, ik zet de verwarming hoger’. Maar voor andere taken is één vraag niet voldoende en gaat de gebruiker **in gesprek** met het apparaat. Het apparaat moet in beide gevallen kunnen spreken – dit noemt men **spraaksynthese**, dan wel *text to speech* (Kuligowska, e.a., 2018). Tijdens zo’n gesprekje wisselen spraakherkenning, interpretatie en spraaksynthese elkaar af.

Samengevat bestaat spraaktechnologie dus uit **spraakherkenning**, **spraakinterpretatie** en **spraaksynthese** (zie figuur 1). Op elk terrein is de afgelopen jaren veel vooruitgang geboekt (zie kader 1). In de volgende paragrafen gaan we nader in op deze drie onderdelen.

Figuur 3 Elementen van spraaktechnologie



Kader 1 Overzicht mijlpalen geschiedenis spraaktechnologie

1922 – **Radio Rex**. Rex was een speelgoedhondje dat uit zijn huisje kwam als je zijn naam noemde. Het hondje reageerde op zijn naam doordat er een ingebouwde elektromagneet in het huisje zat die gevoelig was voor frequenties van ongeveer 500 hertz, het geluid van de klinker “E” in REX.

1939 – **VODER**. Homer Dudley presenteerde VODER (Voice Operating Demonstrator) op de New York World’s Fair. Het was de eerste elektronische variant van eerder ontwikkelde mechanische spraakmachines. Een getrainde medewerker kon door middel van het indrukken van toetsencombinaties spraakgeluiden produceren.

1962 – **Shoebox**. IBM ontwikkelde Shoebox, een computer die zestien gesproken woorden kon herkennen, waaronder de cijfers nul tot en met negen. De Shoebox was gekoppeld aan een rekenmachine en had, zoals de naam doet vermoeden, de grootte van een schoenendoos.

1979 - **MITalk**. Onderzoekers van MIT ontwikkelden MITalk, een systeem dat spraak kon produceren door geluidsgolven samen te voegen. Het werd onder meer gebruikt door Stephen Hawking.

1982 - **Dragon Systems**. Dragon Systems produceerde spraakherkenningssoftware die gebruik maakte van statistische modellen en die gesproken woorden kon omzetten in tekst die op het scherm verscheen. Deze systemen waren de voorlopers van de huidige dicteersoftware.

2001 – **Natural Voices**. AT&T introduceerde in 2001 Natural Voices, een natuurlijk klinkende spraaksynthesizer die kleine stukjes spraak aan elkaar plakte. Het systeem werd veel gebruikt in online applicaties, zoals websites die e-mails hardop konden voorlezen.

2011 – **Siri**. Apple bracht in 2011 Siri op de markt, als softwareonderdeel van de iPhone 4S. Siri luidde de doorbraak in van spraakassistenten die spraak goed genoeg konden herkennen, interpreteren en synthetiseren om een breed scala aan taken uit te voeren.

2.2 Spraakherkenning

2.2.1 Hoe werkt spraakherkenning?

Het doel van spraakherkenning is om een spraakopname om te zetten naar een helder opgestelde tekst (Huang, 2014). Spraaksystemen kunnen dit op twee manieren doen: door een **hybride analyse** uit te voeren, of een **end-to-end analyse**.

Hybride analyse

Een hybride analyse is de meest gebruikte methode om audio om te zetten in tekst, en bestaat uit **vier onderdelen** die samenwerken om het beste resultaat te produceren (Jyothi, 2018).

1. Er vindt een **akoestische analyse** plaats, waarbij de computer het opgenomen geluid in stukjes opdeelt van enkele milliseconden en analyseert (O'Shaughnessy, 2013). Sommige geluidsstukjes zijn luider dan andere en hebben een andere toonhoogte. De stukjes worden dus gewaardeerd op basis van hun volume – in **decibel**, en hun frequentie – in **hertz**.
2. De computer analyseert de stukjes verder met behulp van een **akoestisch model** (Li, e.a., 2017). Dit model onderscheidt kenmerkende eigenschappen van geluid – zogeheten **fonemen**. Denk aan de **b** en de **d** die de woorden **bak** en **dak** van elkaar onderscheiden. Dit zijn de kleinste onderdelen waar taal uit bestaat – en waar je vervolgens dus zinnen van kunt bouwen. De computer berekent op basis van de akoestische analyse welke fonemen waarschijnlijk bij welke geluidsfragmenten horen, en doet vervolgens een voorstel. Bijvoorbeeld dat het geluid bestaat uit **l**, **ay**, en **k**. Deze taak kunnen computers beter uitvoeren door ze met heel veel geluidsfragmenten te trainen (Jyothi, 2017).
3. Het **uitspraakmodel** maakt woorden van de voorgestelde combinaties fonemen: **l**, **ay**, en **k** vormen bijvoorbeeld het Engelse woord **like** (Jyothi, 2017). Deze taak voert de computer doorgaans uit op basis van een woordenboek opgezet door taalexperts dat combinaties van fonemen koppelt aan woorden.
4. Het **taalmodel** maakt ten slotte van de woorden zinnen (Sundermeyer, e.a., 2012). Ook hier berekent de computer wat de kans is dat bepaalde woorden elkaar opvolgen. En ook hier wordt de software getraind op basis van datasets, die bestaan uit vele bestaande teksten met ontelbaar veel zinnen. Het model vraagt eigenlijk: hoe groot is de kans dat bijvoorbeeld het woord 'de hond' gevolgd wordt door bijvoorbeeld 'rent', 'kent' of 'tent'? De

zinsconstructie met de grootste kans komt vervolgens boven. Hier is 'rent' het meest waarschijnlijk, gevolgd door 'kent' - maar in elk geval niet 'tent'.

De verschillende onderdelen worden door een **decoder** onderling vergeleken en herhaald, totdat de computer de meest waarschijnlijke zin produceert (Jyothi, 2017). Vooral bij deze stap is de kansberekening complex, en is dus de meeste rekenkracht vereist.

De afgelopen tien jaar zijn de hybride systemen steeds meer gebruik gaan maken van **deep learning** (Huang, 2014). Dit is een vorm van kunstmatige intelligentie die gebruik maakt van neurale netwerken en grote hoeveelheden trainingsdata. Een neuraal netwerk kun je het beste begrijpen als een complexe rekenmachine, gemodelleerd naar de menselijke hersenen, die met behulp van kansberekening zeer veel verschillende uitkomsten en scenario's kan uitrekenen en vergelijken. Daardoor kan een neuraal netwerk met behulp van trainingsdata zelf leren hoe data het beste gecategoriseerd kunnen worden. De neurale netwerken verbeteren de hierboven genoemde onderdelen die gebaseerd zijn op kansberekening: de akoestische analyse en het taalmodel.²

End-to-end analyse

De meest recente ontwikkeling op het gebied van spraakherkenning is de end-to-end analyse. Bij deze analyse wordt de gehele omzetting van audio naar woorden uitgevoerd door een neuraal netwerk (Chiu, e.a., 2018). Hierbij worden bepaalde resultaten dus niet gecheckt met een taalmodel, maar rekent het model uit welke letters, woorden en zinnen het meest waarschijnlijk passen bij een stukje audio. Het voordeel hiervan is dat deze systemen makkelijk te trainen zijn en minder rekenkracht vereisen dan de hybride systemen (Synnaeve e.a., 2020). Dit zorgt ervoor dat deze systemen makkelijker op apparaten zelf geïnstalleerd kunnen worden – en niet afhankelijk zijn van externe computers die bestanden in de cloud ofwel online opslaan (Li e.a., 2020; Meyer, 2019; Mwit, 2019). Het nadeel van end-to-end systemen is dat ze ook onbegrijpelijke tekst en niet bestaande woorden kunnen registreren, omdat resultaten niet gecontroleerd worden met een door experts aangemaakt taalmodel. Het is daarom vooralsnog niet duidelijk of de end-to-end systemen de hybride systemen zullen vervangen.

2 Alle grote, commerciële spraakherkenningssystemen maken gebruik van neurale netwerken (Kepuska & Bohouta, 2018; Jyothi, 2018).

2.2.2 Hoe goed werkt spraakherkenning?

Bij alle onderdelen van spraakherkenning kunnen fouten ontstaan, wat tot eigenaardige voorbeelden kan leiden. Bijvoorbeeld als de zin 'Recognize Speech Using Common Sense' wordt herkend als de zin 'Wreck a Nice Beach You Sing Calm Incense' (Lieberman, e.a., 2005). In dit geval werden de woorden allemaal onjuist herkend, laat staan dat zij in de juiste volgorde stonden. Soms kunnen hele zinsfragmenten verloren gaan, omdat de microfoon de geluiden slecht ontvangt of omdat er fragmenten verloren gaan in het opknippen van de verschillende audiostukjes. Zo bleek in 2019 de spraakherkenning van het Witte Huis fouten te maken, waardoor een incompleet transcript van een telefoongesprek naar het congres werd gestuurd (Barnes, 2019).

Toch is de spraakherkenning in de afgelopen jaren met sprongen vooruitgegaan (Huang, 2014; Kodish-Wachs e.a., 2018; Meeker, 2018; Sokol, e.a., 2017; Protalinski, 2019). Dit komt met name door de toegenomen **beschikbaarheid van data**, de **toegenomen rekenkracht** en het gebruik van **deep learning-technologie** (Hoy, 2018). Wetenschappers, maar ook de commerciële spraaktoepassingen die al op de markt zijn, hebben de afgelopen decennia steeds meer spraakdata verzameld. Steeds meer taalgebieden worden daarbij ontsloten, zodat de spraaksystemen niet alleen Engels en Spaans, maar ook andere talen kunnen herkennen. Zo kan Google Assistant al Nederlands herkennen. Door de snellere reken capaciteit lukt het om deze grote hoeveelheden data te verwerken en te analyseren. Met behulp van *deep learning* wordt de analyse ten slotte steeds preciezer. Hierdoor wordt spraakherkenning snel beter, en is het voor een groeiend aantal toepassingen succesvol te gebruiken (zie verder hoofdstuk drie).

De vooruitgang van spraakherkenning kun je meten aan de hand van de foutmarge: het percentage verkeerd vertaalde woorden, weggelaten woorden en abusievelijk toegevoegde woorden. De beste spraaksystemen, de spraakassistenten van grote technologiebedrijven zoals Alexa en Google Assistant, bereiken nu een foutmarge van rond de 5%. Dit is een significante prestatie, omdat de foutmarge van menselijke spraakherkenning ongeveer even groot is (Jawad, 2017; Meeker, 2018; Chiu, e.a., 2018).

Deze foutmarge van 5% is echter misleidend. De hoge percentages komen namelijk alleen voor bij spraaksystemen die getraind zijn met zeer specifieke datasets en die functioneren onder optimale condities. Adams (persoonlijke communicatie, 15 juni 2020), projectleider van het team dat Alexa ontwikkelde, zegt dan ook dat deze cijfers misleidend zijn. Als de spraak bijvoorbeeld meer geaccentueerd is of als een geluidsopname meer omgevingsgeluid bevat, zullen de meeste systemen hogere foutpercentages kennen.

Spraaksystemen kunnen ook moeite hebben met de specifieke spraak van kinderen, ouderen, of bepaalde dialecten, omdat van die groepen te weinig data beschikbaar zijn. Bedrijven richten zich op afzetmarkten en talen waarvan veel data beschikbaar zijn, zoals Amerikaans Engels en Mandarijn Chinees. Daarnaast hebben spraaksystemen moeite met de specifieke terminologie van vakgebieden als het recht of de gezondheidszorg.

Spraakherkenning functioneert het beste als aan bepaalde voorwaarden is voldaan: (1) het systeem is getraind met fragmenten van alle sprekers die het systeem gebruiken, (2) het systeem kent alle woorden die worden gebruikt en (3) het systeem is getraind in alle mogelijke opnameomstandigheden (O'Shaughnessy, 2008). Het is een gigantische uitdaging om aan deze voorwaarden te voldoen, vanwege de enorme variatie in taal en locaties waar taal gebruikt wordt. Vaak presteren spraakassistenten vooral goed als de gebruiker gedwongen is een beperkte set simpele commando's te gebruiken – voorwaarde 2. Maar dit betekent natuurlijk dat de gebruiker zich moet aanpassen aan het spraaksysteem.

Bij verschillende spraaksystemen en –toepassingen is de foutmarge ruim hoger dan 5%. De foutmarge van het transcriberen van telefoongesprekken ligt rond de 10 tot 12%, van het dicteren van teksten en ondertitelen tussen de 10 en 20%, en van het transcriberen van een meeting met een groep mensen rond de 40 tot 50% (3PlayMedia, 2019; J. Adams, persoonlijke communicatie, 15 juni 2020). Bovendien verwijzen deze cijfers naar systemen die accentloos Engels herkennen – bij andere talen en accenten maken de systemen meer fouten. Ten slotte kan de foutmarge zelfs hoger dan 50% zijn in specifieke gebruikscontexten, zoals een ziekenhuis (Kodish-Wachs e.a., 2018).

Hoewel spraakherkenning duidelijk vooruit gaat, blijft de uitdaging bestaan om de nauwkeurigheid te verhogen in lastige omstandigheden, en zal er veel werk verzet moeten worden om ook accenten en minder voorkomende talen goed te herkennen (O'Shaughnessy, 2008; J. Adams, persoonlijke communicatie, 15 juni 2020). Zeker voor gebruik bij kritische toepassingen, zoals een medische ingreep of de aansturing van een vrachtwagen, is spraakherkenning nog niet goed genoeg.

2.2.3 Hoe werkt spraakinterpretatie?

Als de vorige stap, de spraakherkenning, goed verlopen is, beschikt een apparaat over de juiste tekst. Soms is dat ook het eindpunt, bijvoorbeeld als je dicteert en de computer de spraak opschrijft. Maar in veel gevallen is de toepassing complexer en geeft de tekst een instructie – bijvoorbeeld om een sjaal te kopen. Dan is het zaak om de tekst goed te interpreteren. Dit noemen we **tekstanalyse**, een soort analyse

die nauw samenhangt met **natural language processing** (NLP). Daarnaast kan het audiobestand ook nader worden geanalyseerd, bijvoorbeeld voor de detectie van een emotie of het opsporen van een ziekte. Dit wordt **stemanalyse** genoemd.³ We bespreken ze na elkaar.

Tekstanalyse

Tekstanalyse is gebaseerd op NLP: het geautomatiseerd interpreteren van tekst. NLP is interdisciplinair, en combineert kunstmatige intelligentie met linguïstiek, computerwetenschappen en informatica (Chowdhury, 2003). De toepassingen van NLP zijn velerlei, en breder dan alleen spraaktechnologie. Denk aan de werking van spellingcorrectie, spamfilters, zoekmachines, en het monitoren van sociale media.

Het begrijpen van tekst is één van de lastigste onderdelen van spraaktechnologie – en dat is niet verrassend. Menselijke taal is namelijk zelden precies, maar juist ongestructureerd, dubbelzinnig en contextafhankelijk. Een zin als ‘ik pak het op’ kan bijvoorbeeld verwijzen naar het oppakken van het fysieke object, maar ook naar het accepteren en uitvoeren van een bepaalde taak. Die subtiele variaties zijn voor spraaksystemen een uitdaging, omdat het lastig kan zijn om te bepalen welke mogelijke interpretatie de juiste is (Manning, 2017). Dat is zelfs voor mensen vaak niet duidelijk – we vragen regelmatig aan elkaar om een verzoek te herhalen of te verduidelijken.

Dit probleem is op twee manieren op te lossen: door te eisen dat gebruikers simpele, eenduidige commando's gebruiken of door het spraaksysteem zelf te verbeteren. De tweede oplossing is natuurlijk het meest aantrekkelijk – als een spraaksysteem je makkelijker begrijpt, is het efficiënter en comfortabeler om ermee te werken.

NLP is gericht op dit laatste en probeert taal te begrijpen door verschillende analyses uit te voeren. Deze kunnen ingedeeld worden aan de hand van drie niveaus: een **syntactische analyse**, een **semantische analyse** en een **pragmatische analyse**. We bespreken ze kort:

1. Een **syntactische analyse** kijkt naar de grammatica van een zin. Het breekt de zin op in woorden en interpunctie en herleidt woorden terug tot de stam. Ook wordt gekeken naar voorvoegsels en achtervoegsels zoals ‘on’ en ‘tje’. Vervolgens worden de woorden grammaticaal ingedeeld, bijvoorbeeld door aan te geven welk woord een lidwoord is en welk woord een zelfstandig naamwoord. Zo vormt zich een grammaticaal overzicht van de tekst.

³ Stemanalyse wordt soms als onderdeel van spraakherkenning gezien. Herkenning en interpretatie zijn nauw met elkaar verbonden. Omdat de interpretatie rijker wordt als stemanalyse wordt toegevoegd, bespreken wij de technologie als onderdeel van spraakinterpretatie.

2. Een **semantische analyse** kijkt naar de betekenis van de woorden: beschrijven ze mensen of locaties of producten? Hierbij *labelt* het systeem de woorden en moet het goed letten op de context – een taak die moeilijk is om te trainen. Het geautomatiseerd begrijpen van taal wordt ook wel *natural language understanding* genoemd (NLU).
3. Bij de **pragmatische analyse** wordt de intentie van een tekst achterhaald. Wat is de boodschap? Er zijn grofweg twee intenties die bij spraaktechnologie vaak voorkomen: iets opzoeken en iets uitvoeren. Die intentie blijkt doorgaans uit de woorden die worden gebruikt, zoals 'wat is' bij het opzoeken of 'speel' bij het uitvoeren van een actie. De gebruiker weet vaak welke specifieke woorden hij moet kiezen om de juiste intentie over te brengen. Hoe meer van deze sleutelwoorden de digitale toepassing begrijpt, hoe beter het systeem de intentie kan achterhalen.

Hedendaagse spraaktechnologie maakt vaak gebruik van een combinatie van deze processen om de betekenis van een tekst te achterhalen.

Stemanalyse

Op basis van spraakdata is het mogelijk om allerlei stemanalyses uit te voeren. Dit vindt plaats aan de hand van zogeheten *vocal biomarkers*, biologische kenmerken die op basis van iemands stemgeluid gemeten kunnen worden. Dit soort analyse heeft zowel de interesse van wetenschappers als van het bedrijfsleven. Op verschillende toepassingen door het bedrijfsleven gaan we in hoofdstuk drie dieper in.

Met stemanalyse kun je onder meer proberen om iemands leeftijd, identiteit, geslacht en lengte vast te stellen (Chaudhari & Kagalkar 2012; Müller, 2006; Pisanski, e.a., 2014). Ook wordt er gewerkt aan het detecteren van ziektes, zoals alcoholisme, slaapapneu, longziektes, depressie, PTSD, emotionele stress, angststoornissen, autisme, parkinson, alzheimer en zelfs hersentumoren (Place, e.a., 2017; Sondhi, e.a., 2015; Senseable Intelligence Group, 2019; Marmar, 2019). Het moge duidelijk zijn: met stemanalyse wordt geprobeerd om waardevolle informatie op te halen. Soms proberen bedrijven die waarde gelijk te benutten, bijvoorbeeld door in een callcenter een medewerker feedback te geven op basis van stemanalyse – bijvoorbeeld door advies te geven als het systeem constateert dat iemand zich opwindt (Simonite, 2018).

Als je stemanalyse bij tekstanalyse voegt, wordt de interpretatie van iemands spraak rijker. Mensen begrijpen elkaar immers ook door niet alleen te letten op de inhoud van woorden, maar ook op de intonatie en wijze waarop iemand iets zegt. Er wordt daarom veel van verwacht, zeker door sommige commerciële partijen. Zo

claimt het bedrijf Vocalis Health dat door middel van stemanalyse een persoonlijkheidsprofiel opgesteld kan worden (Chen, 2019).

2.2.4 Hoe goed werkt spraakinterpretatie?

Zijn de computers van vandaag opgewassen tegen de rijkheid en ingewikkeldheid van menselijke taal? Ook op dit vlak wordt de spraaktechnologie steeds beter. We maken weer een onderscheid tussen tekstanalyse en stemanalyse.

Tekstanalyse

Wederom zorgt de toename van data, rekenkracht en de toepassing van deep learning-systemen ervoor dat de tekstanalyse sneller en accurater kan plaatsvinden (Young, e.a. 2018). Maar hoewel de spraakinterpretatie in het algemeen verbetert, gaan sommige ontwikkelingen sneller dan andere.

Zo kan software steeds beter zinnen grammaticaal ontleden, woorden correct labelen en spamtekst identificeren. Maar het blijft bijvoorbeeld lastig om complexere vragen goed te interpreteren, ambigue woorden te snappen, dialogen te voeren en teksten inhoudelijk samen te vatten (MacCartney, 2011; Young, 2018).⁴

Spraakinterpretatie werkt vooral goed als de omgeving en de gebruiker meewerken – net als spraakherkenning. De omgeving kan meewerken door websites bijvoorbeeld *voice first* in te richten, door bepaalde sleutelwoorden te gebruiken en korte, voor de spraakassistent herkenbare stukjes informatie, in de website te verwerken. Een spraaktoepassing kan zo makkelijk informatie vinden (Koksal, 2018). De gebruiker kan meewerken door specifieke, herkenbare woorden te gebruiken en goed te articuleren. Als gebruikers natuurlijker en gevarieerder gaan spreken, en bijvoorbeeld een metafoor gebruiken, raken spraaktoepassingen al snel de weg kwijt.

Dit betekent dat spraaktechnologie in bepaalde toepassingen succesvol gebruikt kan worden, zoals een spraakassistent die de weersvoorspelling opzoekt en muziek afspeelt, terwijl andere toepassingen nog toekomstmuziek zijn, zoals een spraakassistent met wie je een onderhoudend gesprek kan voeren (zie hoofdstuk drie).

4 <https://nlpprogress.com/> is een website die ontwikkelaars gebruiken om de ontwikkeling van NLP onderwerpen bij te houden.

Stemanalyse

En hoe goed is stemanalyse? Deze vraag draait met name om de kwaliteit van de vocal biomarkers: hebben bepaalde ziektes en bepaalde emoties echt onderscheidende stemkenmerken? Om hier achter te komen moet je twee groepen met elkaar vergelijken, en bijvoorbeeld onderzoeken of een groep alzheimerpatiënten een bepaald stemkenmerk heeft dat bij een groep niet-alzheimerpatiënten ontbreekt. Deze groepen mogen in andere opzichten niet van elkaar verschillen – zo moet onder meer de leeftijd bij de twee groepen gelijk zijn.

Dit voorbeeld illustreert direct de uitdaging van stemanalyse. Er is een hoogwaardige dataset voor nodig om de kunstmatige intelligentie te trainen. Veel onderzoekers geven aan dat er opnames nodig zijn van tienduizenden correct geselecteerde proefpersonen voordat voldoende bewezen kan worden dat de algoritmen werken (French, 2019). Bovendien verschilt een gevonden correlatie van een oorzakelijk verband: vaak is onduidelijk waarom de stemkenmerken correleren met een bepaalde ziekte of een bepaalde eigenschap. Verschillende wetenschappers zijn daarom sceptisch over het vermogen van stemanalyses om eigenschappen te ontdekken en profielen op te stellen (Chen, 2019). Voorzichtigheid is geboden.

Bepaalde eigenschappen, zoals identiteit, geslacht of leeftijdscategorie, lijken met spraaktechnologie wel betrouwbaar vast te stellen (Sedaaghi, 2009; Gupta, e.a., 2019; Abdulsatar, e.a., 2019). Bovendien is niet voor elke toepassing dezelfde mate van betrouwbaarheid nodig. Het diagnosticeren van een ziekte zal aan heel andere eisen moeten voldoen dan het vaststellen van een bepaalde muziek- of filmvoorkeur. Vooralsnog is spraaktechnologie ten aanzien van de meeste eigenschappen niet in staat om voldoende betrouwbare stemanalyses te geven (Wenderow, 2019).

2.3 Spraaksynthese

2.3.1 Hoe werkt spraaksynthese?

Er bestaan ten slotte ook digitale spraaktoepassingen die niet alleen tekst herkennen en interpreteren, maar zelf kunnen spreken. Zo kan bijvoorbeeld informatie over een bekend persoon worden voorgelezen, of bevestigt het digitale spraaksysteem een bepaald commando: 'hier is de Spotify playlist van John Mayer'. Ook kan de toepassing vragen om verdere informatie, om meer ingewikkelde vragen te beantwoorden. We leggen kort uit hoe spraaksynthese werkt.

Grofweg zijn er twee vormen van spraaksynthese: **aaneengeschakelde synthese** en **modelsynthese** (Jothilakshmi & Gudivada, 2016):

1. **Aaneengeschakelde synthese** (ook wel *sample-based synthesis* genoemd) maakt gebruik van een database van ingesproken stukjes spraak en knipt deze op in kleine deeltjes (Kishore & Black, 2003). Deze delen worden vervolgens gebruikt om weer andere woorden van te maken. Het woord “impressive” kan bijvoorbeeld samengesteld worden uit eerdere opnames van “imp” van impossible, “pres” van president, en “ive” van detective. Bij deze vorm van spraaksynthese moet je steeds een afweging maken tussen grotere stukjes spraak die natuurlijker klinken, en kleinere, flexibele stukjes spraak, die schokkend en haperend over kunnen komen, maar wel meer woorden kunnen vormen.
2. **Modelsynthese** (ook wel *parametrische syntheses* genoemd) bootst spraak na door akoestische eigenschappen zoals het frequentiespectrum, de intonatie, en de nadruk, te halen uit datasets van spraakfragmenten (Tokuday & Zen, 2015). Bij deze vorm van spraaksynthese wordt een digitale toepassing getraind op basis van spraakopnames met de bijbehorende tekst. Hierdoor ontstaat een model dat deze akoestische eigenschappen kan herproduceren. Vervolgens wordt er nieuwe tekst ingevoerd die het getrainde model dan met de akoestische eigenschappen omzet naar spraak. Het voordeel van deze methode is dat de stem heel flexibel is – op basis van een beperkt aantal ingesproken stukken tekst kan het model de stem generaliseren en zinnen produceren die niet zijn ingesproken. De precisie van de stem wordt beter naarmate er meer spraakdata ingesproken worden. De toepassing van deep learning heeft deze vorm van synthese verder ontwikkeld (Zen, 2018).

2.3.2 Hoe goed werkt spraaksynthese?

Kunnen computers al praten zoals mensen? Of herkennen we de stem van een robot meteen? Voor de ontwikkelaars van spraaktechnologie zijn deze vragen cruciaal. Zoals Sundar Pichai, algemeen directeur van Google LLC, al over hun digitale spraakassistent zei: ‘We want the assistant to be something that is natural and comfortable to talk to. And to do that we need to start with the foundation of the Google Assistant, the voice’. (2018, 16:55)

Volgens wetenschappers zorgt een menselijke computerstem voor minder misverstanden, zorgt het ervoor dat het aangenamer is om mee te werken en schept het zelfs vertrouwen, vertelt psycholinguïst Wagner (Stinson, 2017). Mensen schrijven van nature al vrij snel menselijke eigenschappen toe aan computers, ook

wel bekend als het ‘Eliza effect’ (Weizenbaum, 1966). Spraakassistenten kunnen dit effect vergroten door een menselijke stem te gebruiken. Maar is het ook gelukt om een menselijke computerstem te maken? Zijn digitale spraaksystemen zo goed dat ze niet meer van mensen te onderscheiden zijn – en dus de zogeheten Turing test⁵ halen?

Deze vraag is niet eenvoudig met ‘ja’ of ‘nee’ te beantwoorden. Aan de ene kant zijn computerstemmen doorgaans van menselijke stemmen te onderscheiden. Aan de andere kant wordt computerspraak snel beter. Zo lukt het ontwikkelaars om spraaksoftware beter gebruik te laten maken van pauzes en intonaties, en om de systemen vloeiender te laten spreken (Oord, e.a., 2016). Google Duplex, een service van Google Assistant die reserveringen voor je kan maken, gebruikt klanken als “uhm” en “hmm” en informeel taalgebruik (“oh yeah”) om nog menselijker over te komen. De technologie achter Duplex produceert spraak die soms niet van menselijke spraak te onderscheiden is, zoals Sundar Pichai demonstreerde tijdens de Google ontwikkelaarsconferentie van 2018.⁶

Het is overigens niet verrassend dat dit bij Duplex is gelukt. De toepassing beperkt zich namelijk tot een enkelvoudige taak: het maken van reserveringen. Hierdoor kan er veel aandacht besteed worden aan het verbeteren van een kleine set antwoorden. Wat ook meespeelt, is het feit dat Duplex een Engelse stem heeft. Voor het spreken van de Engelse taal zijn meer dan genoeg data voorhanden. De vooruitgang blijft immers afhankelijk van de beschikbare data. Tegelijkertijd is duidelijk dat als de data eenmaal verzameld zijn, spraaksynthese van hoog niveau bereikt kan worden.

Maar ook buiten meer specifieke domeinen verbetert spraaksynthese. Recente ontwikkelingen laten zien dat het, ongeacht de toepassing, ook met relatief kleine datasets lukt om een stem te genereren (Arik e.a., 2018). Zo heeft het bedrijf Resemble AI een opname van minimaal drie minuten nodig om een stemsynthese te maken (Schwartz, 2019). In 2017 liet de startup Lyrebird zien dat ze met een geluidsfragment van zestig seconden een stemsynthese konden maken. Ze demonstreerden dit door een audiofragment te presenteren waarin stemkarakteristieken van Barack Obama, Donald Trump en Hilary Clinton te horen waren (Vincent, 2017). Vooralsnog waren deze fragmenten nog wel duidelijk als computergegenereerd te herkennen.

⁵ Turing verwijst naar Alan Turing, een Britse wiskundige, computerpionier en informaticus.

⁶ Ook andere systemen van Google behalen op het gebied van spraaksynthese een hoge kwaliteit. De kwaliteit van de stemsynthese wordt gemeten via de Mean Opinion Score, die gebaseerd is op de beoordelingen van een groep mensen (Streijl, Winkler, & Hands, 2014). Zij geven dan een score tussen de 1 en 5. WaveNet, een spraaksysteem van Google, scoort hierbij bijvoorbeeld een 4.21 voor Amerikaans Engels en 4.08 voor Mandarijn Chinees (Oord, 2016).

Met behulp van zogeheten “*voice cloning*” kunnen spraaktoepassingen gepersonaliseerd worden. In plaats van een voorgeprogrammeerde stem te kiezen, kan een gebruiker bijvoorbeeld zijn eigen stem instellen, of die van familie of vrienden. Het navigatiesysteem van Waze biedt deze dienst al aan (Welch, 2017). Om hun spraakassistenten nog aantrekkelijker te maken, breiden de grote technologiebedrijven hun stemaanbod nu ook uit met beroemdheden zoals Samuel Jackson (Amazons Alexa) en John Legend (Google Assistant) (Faulkner, 2019).

Uiteindelijk is het doel van ontwikkelaars om een zo divers mogelijk aanbod van stemmen te creëren, waardoor gebruikers, maar ook bedrijven en organisaties, de stem kunnen kiezen die het beste past bij hun identiteit of merk. Dit wordt ook wel *voice branding* genoemd.

Samengevat produceren vrijwel alle spraaksystemen spraak die goed verstaanbaar is, en kunnen sommige systemen ook complexe zinnen formuleren. Voor steeds meer praktische toepassingen, van een navigatiesysteem tot een reserveringsdienst, is het spraakvermogen voldoende. Ook werkt men aan het klonen van stemmen. De technologische ontwikkeling richt zich nu vooral op het menselijker maken van de stemmen – daar kan nog wel de nodige vooruitgang worden geboekt.

2.4 De toekomst van spraaktechnologie

Dit hoofdstuk gaf een overzicht van de huidige mogelijkheden van spraaktechnologie. Maar wat mogen we in de toekomst verwachten? We zetten de ambities van de ontwikkelaars op een rij.

Ten aanzien van **spraakherkenning** willen ontwikkelaars meer talen ontsluiten – op dit moment werkt spraakherkenning vooral goed voor veel gesproken talen zoals Amerikaans Engels of Chinees Mandarijns. Zo is er op dit moment maar één spraakassistent die het Nederlands voldoende beheerst: Google Assistant. Deze ontwikkeling hangt vooral samen met de uitdaging om genoeg data te verzamelen van kleinere taalgebieden. En met de bereidheid van ontwikkelaars om producten te maken voor een kleinere afzetmarkt.

Ook willen ontwikkelaars graag de foutmarge naar beneden brengen. Dit moet gebeuren door de digitale modellen zodanig te verbeteren, dat de software ook dubbelzinnige, zeldzame of vage woorden en uitdrukkingen gaat begrijpen. Daarnaast is het zaak ruis te reduceren, en te werken aan spraaktechnologie die ook goed werkt in lastigere omgevingen zoals een drukke stationshal of zelfs een dansfeest.

De toekomst van **spraakinterpretatie** lijkt vooral te liggen in het combineren van verschillende betekenisvolle elementen, zoals de woorden die iemand gebruikt, de manier waarop iemand de woorden spreekt, en de lichaamstaal, gezichtsuitdrukking en emotie waarmee de woorden worden gesproken (Jyothi, 2018; Henzi & Wright, 2019; Mattin, 2019) (zie ook hoofdstuk drie). Ook is het de bedoeling dat spraaksystemen nog veel beter de context van een vraag of opmerking gaan begrijpen, en het doel van een gesprek helder voor ogen houden (Hirschberg & Manning, 2015; Pundak, e.a. 2018; Zen, 2018).

Bij het **synthetiseren van spraak** willen ontwikkelaars in de toekomst de gebruiker nog centraler stellen – hij of zij moet het gevoel hebben dat er op een persoonlijke en vriendelijke manier met hem of haar gepraat wordt, en dat de ontvangen informatie precies aansluit op zijn of haar behoeftes (Pichai, 2019). Dit zal zonder twijfel persoonlijke gegevens vergen. Het is overigens de vraag of we wel willen spreken met computers die niet van mensen te onderscheiden zijn, en of we alle benodigde persoonlijke informatie wel willen overdragen – hier gaan we in hoofdstuk vier verder op in.

Bij elkaar opgeteld willen ontwikkelaars spraaktechnologie maken die in elke denkbare ruimte te gebruiken is, accuraat tal van diensten uitvoert en velerlei vragen beantwoordt, en samenwerkt als een invoelende, persoonlijk assistent. Dit zal om meer data, meer intieme persoonlijke data en slimmere digitale modellen vragen. Bovendien zal het nodig zijn om de benodigde rekenkracht behapbaar te houden of zelfs te verkleinen – dan kan spraaktechnologie makkelijker in velerlei apparaten geplaatst worden (Pichai, 2019; Li, e.a. 2020).

2.5 Conclusie

In dit hoofdstuk onderzochten we wat spraaktechnologie precies is en hoe goed deze technologie werkt. Bij elk centraal onderdeel van spraaktechnologie – spraakherkenning, –interpretatie en –synthese –, is het afgelopen decennium indrukwekkende vooruitgang geboekt. Het is inmiddels mogelijk om vragen te stellen aan een digitaal apparaat en direct een goed verstaanbaar antwoord te krijgen. Die vooruitgang komt door de enorme toename van beschikbare data om de software mee te trainen, verbeterde digitale modellen en toegenomen rekenkracht. Met name de inzet van *deep learning* heeft een snelle verbetering van spraaktechnologie teweeg gebracht.

Tegelijkertijd is het duidelijk dat de drie onderdelen van spraaktechnologie ook hun beperkingen kennen. Er is doorgaans hulp van de omgeving en van de gebruiker

nodig om spraaktechnologie goed te laten functioneren. Op het gebied van spraakherkenning wordt de mate waarin systemen fouten maken kleiner, maar zijn de foutmarges nog steeds significant en in lastige situaties zelfs hoog. Voor kritische toepassingen is de technologie nog niet goed genoeg. Ontwikkelaars verbeteren daarom niet alleen hun systemen, maar willen ook graag dat de digitale omgeving *voice first* wordt ingericht, door herkenbare sleutelwoorden en korte stukken tekst die goed te vinden zijn door spraaktechnologie.

Bij spraakinterpretatie is ook vaak de hulp van de gebruiker nog nodig om de intentie van gegeven commando's, zeker als ze ingewikkeld zijn, helder te krijgen. Hoewel een eenvoudige opdracht uitvoeren goed lukt, blijven ingewikkeldere conversaties lastig. Om de interpretatie te verbeteren, willen ontwikkelaars verschillende soorten gegevens, waaronder de gesproken woorden en de emotie waarmee iemand iets zegt, met elkaar combineren. Op het gebied van stemanalyse zijn de ambities van technologie- en dienstaanbieders groot, maar wetenschappers zijn vooralsnog kritisch over de betrouwbaarheid van dit soort toepassingen.

Ten slotte ligt de uitdaging van spraaksynthese vooral op het gebied van het verder personaliseren van het stemaanbod en het nog menselijker maken van de stem. Ontwikkelaars willen uiteindelijk beschikken over systemen waarvan de spraak niet meer van menselijke spraak te onderscheiden is, en die een rijk scala aan stemmen en persoonlijkheden kunnen synthetiseren.

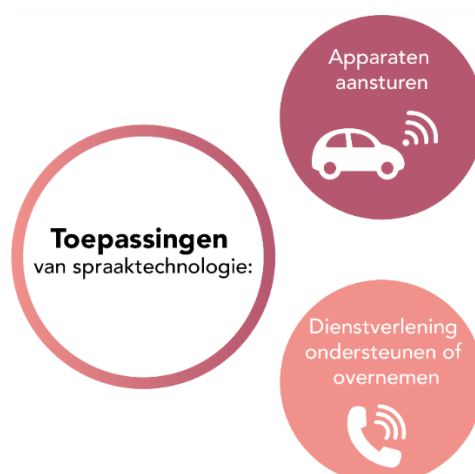
3 Waar wordt spraaktechnologie voor gebruikt?

Dit hoofdstuk geeft een overzicht van het huidige gebruik van spraaktechnologie in de samenleving. Het bespreekt **de toepassingen** die spraaktechnologie kan uitvoeren (3.1). Vervolgens kijken we naar **de analyses** die op basis van spraakdata te maken zijn (3.2) en geven we een beeld van **de markt** voor spraaktechnologie (3.3). Het hoofdstuk maakt met name gebruik van het werk van onderzoeksbureaus en wetenschappers die de markt voor spraaktechnologie in kaart brengen.

3.1 De toepassingen die spraaktechnologie kan uitvoeren

Spraaktechnologie kan uiteenlopende toepassingen uitvoeren. We geven een overzicht van de domeinen waar spraaktechnologie geïntroduceerd wordt, op basis van twee categorieën: spraaktechnologie die **apparaten aanstuurt**, en spraaktechnologie die **dienstverlening ondersteunt of overneemt** (zie figuur 2). In sommige domeinen speelt spraaktechnologie al een veel grotere rol dan in andere. Zo kent de markt voor spraaktechnologie in huis en in de auto al veel gebruikers, terwijl in andere domeinen nog vooral met spraaktechnologie wordt geëxperimenteerd, en onduidelijk is hoe de technologie uiteindelijk het meeste ingezet zal worden. We beschrijven eerst de toepassingen in huis en in de auto.

Figuur 4 Toepassingen van spraaktechnologie



3.1.1 Apparaten aansturen

De afgelopen jaren zijn vele apparaten met het internet verbonden, en is een *Internet of Things* ontstaan (Vermeend & Timmer, 2016). Dit netwerk kan in toenemende mate met spraaktechnologie bestuurd worden. We bespreken vier domeinen waar spraaktechnologie een rol speelt in de aansturing van apparaten: **thuis**, in de **auto**, in de **industrie** en in **wearables** (zie figuur 3).

Figuur 5 Aansturing met spraaktechnologie



Thuis

Onze huizen zijn al vaak zogeheten *smart homes*, waarin huishoudelijke elektronica met elkaar verbonden zijn. Van vrijwel elk elektronisch apparaat is een 'slimme' variant beschikbaar die aan het internet te koppelen is: de tv, lichten, thermostaten, videodeurbellen, camera's, huishoudelijke apparaten, en natuurlijk de smart speaker. In Nederland hebben 3,3 miljoen huishoudens één of meer smart home-producten waarvan de meeste vallen onder de categorieën 'verlichting & schakelaars' en 'energie & verwarming'. Het aandeel van huishoudens dat een smart speaker in huis heeft staan groeit het snelst: van 6% in 2018, tot 19% in 2019 (Multiscope, 2020).

Deze apparaten kun je koppelen aan applicaties op een smartphone of tablet – en die applicaties kun je vervolgens via een spraakassistent aansturen. Dit betekent dat iemand, zittend aan zijn eettafel, met zijn stem de tafellamp kan dimmen, de verwarming hoger kan zetten of bepaalde muziek af kan spelen. Hij kan zelfs een bepaalde 'routine' instellen, en simpelweg zeggen 'Hey Google, etenstijd'. De assistent kan dan verschillende acties combineren, en tegelijkertijd het licht zachter zetten en zorgen voor sfeervolle muziek.

De toepassing van spraak in de *smart home* is niet alleen een kwestie van comfort. Wetenschappers, bedrijven en (zorg)instellingen kijken al decennia hoe slimme apparaten ouderen of mensen met een beperking langer zelfstandiger kunnen laten wonen (bijvoorbeeld onder de term domotica, of *Ambient Assisted Living*). Ook spraak biedt hiertoe mogelijkheden. Ouderenbelangenorganisatie ANBO, verzekeraar Achmea, de Sociale Verzekeringsbank en Google hebben daarom een proef opgezet waarbij ouderen een Google Home in huis kregen (Lens-FitzGerald e.a., 2020).

Ten slotte bevat slim speelgoed vaak spraaktechnologie. Zo is er een miljardenmarkt ontstaan op het gebied van geautomatiseerd speelgoed, dat allerlei taken kan uitvoeren, van het afspelen van muziek tot het voeren van gesprekjes. Bijvoorbeeld de robothond Golden Pup, die met specifieke blafjes reageert op de spraak van een kind. Of Hello Barbie, een barbiepop die kan terugpraten, allerlei spelletjes kan spelen, en kinderen zelfs aanleert om 'sorry' te zeggen (Vlahos, 2015). Naast fysiek speelgoed zijn er ook spraakgames, zoals Wild Island, die je op je telefoon of op een slimme speaker kan spelen. Daarbij wordt bijvoorbeeld een interactief verhaal verteld, waarbij de luisteraar keuzes moet maken.

In de auto

Een andere omgeving waar spraaktechnologie al langere tijd gebruikt wordt, is de auto, met name om *hands free* te kunnen bellen. De laatste tijd worden meer functies geïntroduceerd, zodat de chauffeur zijn handen aan het wiel houdt en zijn aandacht op de weg kan richten (Muller, 2019). Andere functies zijn bijvoorbeeld navigeren, tekstberichten verzenden, en audio afspelen zoals een radiozender of een podcast, die met name in de Verenigde Staten met regelmatig worden gebruikt.⁷ In Nederland is hier nog minder over bekend. Alexa is nog niet beschikbaar in het Nederlands, en ook Android Auto (Google Assistant) is nog niet officieel beschikbaar (Kamp, 2019).

7 In de Verenigde Staten zijn er 77 miljoen maandelijkse gebruikers, meer dan het gebruik van smart speakers in huis, dat op 45.7 miljoen maandelijkse gebruikers staat (Kinsella & Mutchler, 2019).

Verscheidende automerken, waaronder Mercedes en BMW, ontwikkelen zelf een assistent en dus ook een stem voor in de auto. Daarbij wordt naar de stem gezocht die het merk het beste kan ondersteunen. Zo kan voor een jonge, vrouwelijke en intelligente stem gekozen worden, om zo te proberen een bepaald publiek aan te spreken. In hoofdstuk vier bespreken we verschillende ethische aspecten die bij het selecteren van stemprofielen een rol spelen. Marketingkeuzes ten aanzien van stemprofielen vallen in het nieuwe domein van *voice branding* (Vernuccio e.a., 2020).

Spraak kan ook gebruikt worden om onderdelen van de auto zelf aan te sturen, zoals het aanzetten van de verwarming of het openen van de ramen, via de auto of via smartphones (Carmen, 2018). Verschillende patenten van bedrijven laten zien dat er gewerkt wordt aan het aansturen van zelfrijdende auto's via spraaktechnologie (Szymkowski, 2020). Ook wordt er gekeken naar het aansturen van tractoren of graafmachines met spraaktechnologie. Vanwege een te grote foutmarge lijkt de technologie hiervoor voorlopig nog niet geschikt (Kartalidis, 2018).

In de industrie

Spraaktechnologie biedt mogelijkheden om productieprocessen te verbeteren. Een bekende toepassing in de logistieke sector is *Pick by Voice*. In het distributiecentrum worden medewerkers via spraak aangestuurd. In het oor krijgen zij te horen op welke plek ze iets moeten ophalen, hoeveel ze moeten ophalen en hoe ze het meest efficiënt door een distributiecentrum kunnen bewegen (Glynn, 2020).

Nieuwe toepassingen zijn spraakassistenten in de fabriekshal en maakindustrie. Zo bouwde het bedrijf IT Speex de spraakassistent Athena, speciaal ontwikkeld om op de fabrieksvloer machines aan te sturen, zoals boormachines en draaibanken (Leventon, 2019). De assistent kan samen met een scherm gebruikt worden. Het is de bedoeling dat de spraakassistent niet alleen ondersteuning biedt bij de directe aansturing van apparaten, maar ook bij onderhoudschecks, het plannen van productieprocessen en het berekenen van kosten – het moet een assistent zijn die bij vrijwel al het denkwerk en aansturingswerk van waarde is. Net als in de auto en bij graafmachines, is een lage foutmarge essentieel. Zo zijn de spraakherkenningscores van de huidige spraakassistenten, die in optimale omstandigheden uitkomen op 5%, te laag voor industriële toepassingen.

Spraakontwikkelaars zoeken daarom naar manieren om de foutmarge te verlagen. Bovendien werkt een assistent als Athena met expliciete toestemming: als de assistent van plan is een taak uit te voeren, wordt dat eerst aangekondigd. De gebruiker geeft vervolgens toestemming. Daarnaast legt de assistent bij

onduidelijkheid de bal terug bij de gebruiker – bij onduidelijke woorden wordt er niet gegokt naar de betekenis ervan. Ten slotte kan een spraakassistent ook de veiligheid verhogen door bij riskante processen checklists af te gaan en de gebruiker vragen te stellen.

In wearables

Spraaktechnologie is steeds vaker draagbare technologie, die op **wearables** zoals slimme horloges, oortjes en brillen is geïnstalleerd (Williams, 2019). Horloges en oortjes zijn voor de consument beschikbaar, terwijl slimme brillen vooralsnog een focus hebben op sectoren zoals de bouw, de zorg en de logistiek. Slimme brillen kunnen informatie weergeven in het gezichtsveld en geluid afspelen via ingebouwde speakers of soms zelfs door beengleiding.⁸ Een voorbeeld hiervan is de Google Glass die met behulp van spraak en een touchpad te besturen is. Het navigeren door het menu, het starten van een video-opname, en het bellen met vrienden of collega's verloopt bijvoorbeeld via spraak. Slimme brillen worden met name gebruikt in de professionele sector om werkprocessen te verbeteren. Zo laat het aankomende rapport van het Rathenau Instituut over *augmented reality* zien hoe spraaktechnologie wordt gebruikt in de bouw, logistiek en neurochirurgie (zie Sniijders e.a., 2020). Virtual reality headsets scharen we hier niet onder wearables, omdat je hiermee afgesloten bent van de fysieke omgeving. Wel experimenteren ontwikkelaars van deze headsets met spraakaansturing (Zhu e.a., 2015).

3.1.2 Dienstverlening ondersteunen, of overnemen

Spraaktechnologie kan op talrijke manieren het verlenen van diensten ondersteunen, en soms zelfs overnemen. Zo kan een spraakassistent, net als een menselijke assistent, mensen helpen om hun agenda bij te houden, afspraken in te plannen, een to-dolijst bij te houden en berichten te versturen. De assistent van Microsoft, Cortana, richt zich in het bijzonder op dit soort algemene werkprocessen. De assistent kan het werk in eigenlijk alle domeinen versnellen en zo tijd besparen. In de meeste domeinen wordt met spraaktechnologie geëxperimenteerd, en wordt de technologie nog niet breed gebruikt.

Er worden ook talrijke spraaksystemen gemaakt voor meer specifieke dienstverlening. We geven hier een breed, maar niet uitputtend overzicht, en bespreken kort toepassingen op het gebied van **reizen** en **toerisme**, **winkelen**, de **zorg**, het **onderwijs**, **callcenters**, **media-uitzendingen**, **taalwerk**, **identificatie** en **verificatie**, **online zoeken** en **advertenties** (zie figuur 4).

8 Beengleiding werkt door geluid door te geven via trillingen door het bot van de schedel.

Figuur 6 Dienstverlening met spraaktechnologie



Bedrijven werken aan spraakassistenten op het gebied van **reizen** en in de **toerismesector**. Bijvoorbeeld om restaurants te reserveren, zoals Google Duplex doet. Alexa for Hospitality, een versie van Alexa die Amazon speciaal heeft ontwikkeld voor hotels, helpt bij het bestellen van roomservice of extra handdoeken, het aansturen van de airco en het uitchecken (Amazon, 2020). Ook in de luchtvaart kunnen spraakassistenten al taken verrichten. Zo kan de 'slimme assistent' van de Koninklijke Luchtvaart Maatschappij (KLM) je helpen met het vinden van een bestemming, het zoeken naar een vlucht en het opstellen van een paklijst, en je vertellen hoe laat je naar Schiphol moet vertrekken. In de bancaire sector vind je vergelijkbare dienstverlening. Zo kun je de slimme assistent van de Rabobank vragen naar je saldo, verzoeken om een betaling te doen of een overzicht laten geven van je laatste transacties.

Winkelen is ook een opkomend domein voor spraaktechnologie. De toepassingen bestaan deels uit het direct online bestellen van producten en diensten. In de

Verenigde Staten kun je de spraakassistent Alexa bijvoorbeeld al vragen om huishoudelijke producten bij te bestellen: 'Alexa, order more dog food.' Dit staat in Nederland nog in de kinderschoenen – met name omdat de grootste Nederlandssprekende assistent, Google Assistant, nog geen bestellingen kan doen. Google werkt wel aan een betaalfunctie voor Google Assistant (Hager, 2020).

Deels bestaat de spraaktechnologie uit het ondersteunen van bestellingen, bijvoorbeeld door keuzes aan te raden en boodschappenlijstjes bij te houden. Dit komt in Nederland al meer voor. Zo heeft Albert Heijn een spraakapplicatie gemaakt waar je aan kunt vragen hoe laat de bezorger komt, welke producten in de bonus zijn of welke recepten er zijn voor een ovenschotel met groenten. Albert Heijn wil zo boodschappen doen en koken vergemakkelijken (Van Bergen & Van Gelder, 2018). Bol.com en Jumbo zijn ook aan de slag gegaan met spraaktechnologie die geen bestellingen doet, maar daar wel bij helpt.

Amazon ziet in spraak 'een nieuw tijdperk in de detailhandel' (Meridith, 2019). Critici wijzen er echter op dat maar 2% van de Alexagebruikers ooit producten heeft gekocht door middel van spraaktechnologie (Pardes, 2019). Slechts 10% van deze gebruikers deed dit daarna nog een keer (Anand, 2018). De belangrijkste reden hiervoor is dat winkelen in veel gevallen een visuele activiteit is. Je wilt zien wat je koopt, en producten overzichtelijk met elkaar vergelijken. Daar zijn schermen geschikter voor dan smartspeakers – dus misschien is de combinatie van spraak en scherm wel de toekomst van commerciële transacties.

Een ander opkomend toepassingsterrein is de **zorg**. Er wordt spraaktechnologie ontwikkeld om spraakassistenten te gebruiken als eerste loket: je kunt bij de assistent een afspraak maken, recepten aanvragen en de assistent poogt door middel van vragen een basale triage te verrichten, en je vervolgens door te verwijzen naar de juiste zorgverlener. Ook werken leveranciers van het Elektronisch Patiëntendossier samen met Amerikaanse bedrijven om door middel van spraaktechnologie patiëntendossiers bij te werken, verwijsbrieven te schrijven en nota's op te stellen (Jacobs, 2018). Deze innovaties worden gezien als een kans om de zorg efficiënter te maken, en de groeiende zorgkosten te drukken. Bovendien proberen de ontwikkelaars de spraakassistent zo invoelend en empathisch mogelijk te maken (Kritzler, e.a., 2019) De spraakassistent kan de tijd nemen voor patiënten, en ze geduldig de weg wijzen in het complexe en drukke zorgdomein.

Naast het vergemakkelijken van administratieve processen, dienen spraakassistenten ook het werk van artsen te ondersteunen. Wolters Kluwer (2020) werkt bijvoorbeeld samen met spraakbedrijf Nuance om artsen klinische databases via spraak te laten raadplegen. Alexa biedt een geautomatiseerde, medische

transcriptiedienst voor medische professionals. Wetenschappers en bedrijven richten zich ook op het diagnosticeren van ziekten op basis van stemdata (zie paragraaf 3.3).

Dit geldt ook voor het **onderwijs**. Zo heeft een consortium van onderzoeksinstellingen en commerciële partijen iTalk2Learn ontwikkeld, een open source onderwijsplatform waarmee leerlingen wiskunde kunnen leren (iTalk2Learn 2015). De software stelt vragen en herkent de spraak van leerlingen, en analyseert hun houding – valt een bepaalde opdracht in de smaak of juist niet? Ook Duolingo, een veel gebruikte applicatie waarmee je vreemde talen kunt leren, maakt gebruik van spraaktechnologie, met name spraakherkenning (Sawers, 2019). Er wordt gewerkt aan bots waarmee gebruikers kunnen converseren.

In **callcenters** wordt ook ingezet op spraaktechnologie. Zo ondersteunt het bedrijf Cogito medewerkers met behulp van spraakherkenning. Het detecteert tijdens gesprekken of de beller enthousiast of gefrustreerd is, en geeft vervolgens real time advies aan de medewerker. De medewerker kan ook een notificatie krijgen om een pauze of kopje koffie te nemen wanneer zijn of haar stem vermoeid klinkt (Dzieza, 2020). Er worden ook spraakassistenten aangeboden zoals Amelia, van ontwikkelaar IPsoft, die zelf het gesprek aangaan om vragen die bellers hebben te beantwoorden (Baraniuk, 2018). Bovendien helpen de eerdergenoemde assistenten van bedrijven zoals Albert Heijn en KLM ook bij het beantwoorden van algemene vragen van klanten.

Spraaktechnologie wordt ook ingezet bij **media-uitzendingen**. Dit gebeurt met name door spraakassistenten te vragen om bepaalde programma's af te spelen, zoals het nieuws: ze spelen dan een ingesproken nieuwsbericht af dat is klaargezet door een bepaalde journalistieke bron, zoals een krant of het NOS journaal. De gebruiker kan de bron instellen. De Nederlandse Publieke Omroep heeft uitvoerig samengewerkt met Google om te onderzoeken welke NPO-programmering het beste via spraakassistenten en smartspeakers kan worden aangeboden (NPO, 2020). In China gaat men al een stap verder: daar heeft Xinhua News Agency, een persagentschap van de Chinese staat, een mannelijke en een vrouwelijke AI-nieuwslezer gebouwd, gebaseerd op menselijke nieuwslezers. Deze zijn al ingezet tijdens grote evenementen in China (Meiling, 2019).

Spraaktechnologie kan ook allerlei **taalwerk** overnemen. Denk bijvoorbeeld aan het vertalen van spraak. Tijdens een gesprek tussen twee personen die elkaars taal niet spreken, kan spraaktechnologie zoals Google Translate en Skype translator, onmiddellijk zinnen van de ene persoon herkennen en vertalen (Kohn, 2019). Spraaktechnologie kan ook vergaderingen notuleren en opnames transcriberen. Ook is het mogelijk om met spraaktechnologie filmpjes of live presentaties

geautomatiseerd van ondertiteling te voorzien. Al deze toepassingen zouden een grote potentie kunnen hebben: stel je maar een wereld voor waar iedereen met een smartphone beschikt over vertaaltechnologie, en dus relatief eenvoudig communiceert met mensen die een geheel andere taal spreken. Deze toepassingen vereisen wel dat spraaktechnologie veel meer talen beheerst dan nu het geval is.

Ook in het veiligheidsdomein wordt spraaktechnologie als veelbelovend gezien. Met behulp van spraaktechnologie kunnen mensen namelijk **geïdentificeerd** en **geverifieerd** worden. Bij identificatie vergelijk je iemands stemgeluid met een brede dataset, en achterhaal je zo iemands identiteit. Hiervoor is een uitvoerige dataset vereist. Verificatie werkt anders. Daarbij claimt iemand een bepaalde identiteit en stelt de software op basis van een geluidsopname vast dat een stem authentiek is, en bij de opgeslagen identiteit hoort. Hiervoor is slechts een vooraf opgeslagen opname per persoon nodig. Deze keuze is van belang voor de impact van spraakauthenticatie op privacy en beveiliging van gegevens (zie hoofdstuk vier).

Banken en telecombedrijven verkennen de mogelijkheden van spraakverificatie (Bharadwaj, 2019). Bijvoorbeeld om klanten die naar de klantenservice bellen te identificeren, of om betalingen te doen of saldo op te vragen. Ook verschillende overheidsinstellingen zijn met spraakverificatie aan de slag gegaan. De belastingdienst van Nieuw-Zeeland maakt bijvoorbeeld gebruik van deze technologie. Burgers en bedrijven konden informatie opvragen en gegevens wijzigen, nadat hun spraak was geverifieerd (Sankaranarayanan, 2017).

Spraakverificatie kan een extra veiligheidsfactor bieden om online fraude tegen te gaan of gebouwen te beveiligen. In sommige gevallen kan het gebruikt worden als een handtekening. In de Verenigde Staten kunnen spraakopnames met de woorden 'I agree' bijvoorbeeld gebruikt worden om een digitaal contract te ondertekenen.

Een stem-ID kan ook gebruikt worden voor het opsporen van criminelen. Dan gaat het om identificatie. Zo kan een geluidsfragment tijdens een winkelberoving worden beluisterd, en kan de politie de daders op basis van een brede dataset van spraak-IDs identificeren. Interpol maakte in 2018 bekend dat het een internationale spraakidentificatiedatabase had opgebouwd, met stem-IDs van 192 verschillende wetshandavingsinstanties van over de hele wereld (Kofman, 2018).⁹

Ten slotte kan spraaktechnologie de manier waarop we **online zoeken** veranderen, en daarmee de manier waarop we worden blootgesteld aan **online advertenties**.

9 <https://www.interpol.int/en/Who-we-are/Legal-framework/Information-communications-and-technology-ICT-law-projects/Speaker-Identification-Integrated-Project-SIIP>.

Door spraaktechnologie kunnen we namelijk zoeken via een spraakassistent, die ons doorgaans niet, (zoals een online zoekmachine), een lijst geeft met zoekresultaten, maar een antwoord uitspreekt – of een vervolgvraag stelt. Deze manier van zoeken heeft grote gevolgen. Allereerst voor partijen die op het internet gevonden willen worden. Zij moeten hun websites *voice first* maken, zodat de spraakassistent de informatie makkelijk vindt. Hier zijn bijvoorbeeld verschillende overheidsorganisaties, zoals de Dienst Uitvoering Onderwijs (DUO), de Kamer van Koophandel (KvK), het Uitvoeringsinstituut Werknemersverzekeringen (UWV), en de Sociale Verzekerings Bank (SVB) al mee bezig.

Daarnaast zal de online advertentiemarkt door het zoeken via spraaktechnologie veranderen (Olson & Kemery, 2019). Je kunt namelijk geen advertentie meer plaatsen boven een online zoekpagina of proberen om ergens in een lijst met zoekresultaten te staan. Hoewel er waarschijnlijk met nieuwe vormen van advertenties geëxperimenteerd zal worden, is er in de omgeving van spraakassistenten vooralsnog minder advertentieruimte. Bedrijven en organisaties zullen er dus voor moeten zorgen dat spraakassistenten hun websites makkelijk kunnen vinden, en dat hun producten en diensten dermate bij gebruikers in de smaak vallen, dat de spraakassistenten ze gaan aanbevelen.

3.2 Data-analyse door bedrijven

Bij de uitvoering van alle besproken taken en rollen verzamelen en verwerken spraaktoepassingen data. In deze paragraaf geven we een overzicht van de manieren waarop bedrijven deze nieuwe bronnen van data analyseren. Daarbij maken we, net als in hoofdstuk twee, een onderscheid tussen de analyse van de **tekst** en de analyse van de **stem**.

3.2.1 Tekstanalyse

De gespreksverslagen van spraaktoepassingen geven weer wat je zegt tegen een spraaksysteem, zoals de woordkeuze en zinsopbouw van een bepaalde gebruiker. De grote technologiebedrijven als Google, Apple en Amazon analyseren de gespreksverslagen om hun spraakassistenten Google Assistant, Siri en Alexa te personaliseren. Zodat de assistent zich bijvoorbeeld kan aanpassen aan je spraakpatronen, vocabulaire en persoonlijke voorkeuren, en zo bepaalde diensten of taken – zogeheten ‘skills’ – aanbeveelt (Dawar, 2018). De techbedrijven passen daarbij emotieherkenning toe, waarbij de emotionele gemoedstoestand van de gebruiker achterhaald wordt. Als Alexa immers weet wat de gebruiker prettig vindt

en wat niet, helpt dat enorm bij de personalisatie van de spraakassistent, en is het waarschijnlijker dat de gebruiker daarmee in gesprek blijft. Emotieherkenning is voor empathische, intelligente spraakassistenten een sleuteltechnologie.

3.2.2 Stemanalyse

Ook de audio-opnames worden gebruikt om de assistent zich te laten aanpassen aan de gebruiker. Deze analyse is bijvoorbeeld gericht op het herkennen van volume. Zo kan Alexa terugfluisteren als je 's ochtends vroeg iets zachtjes vraagt en je wilt voorkomen dat je partner wakker wordt (Gershgorn, 2019).

Uit verschillende patentaanvragen en toepassingen blijkt dat bedrijven ook andere analyses op basis van audio verkennen, met name emotieherkenning (Fussel, 2018). Zo heeft Amazon een toepassing ontwikkeld voor het herkennen van frustratie, zodat ze Alexa zo kunnen ontwikkelen dat de assistent zo min mogelijk frustratie oproept (Gershgorn, 2019). En het bedrijf Cogito monitort met stemanalyse gesprekken met de klantenservice, en geeft medewerkers suggesties mee zoals 'praat rustiger' of 'bied je excuses aan' (Simonite, 2018).

Bedrijven interesseren zich ook voor gezondheidsanalyses – het herkennen van ziektes (Kinsella, 2018). Denk bijvoorbeeld aan de spraakassistent Alexa, die kippensoep suggereert als ze hoort dat de gebruiker hoest. Ook de start-up Vocalis Health doet behulp van stemanalyse gezondheidsonderzoek. Wenderow (2020), de algemeen directeur van Vocalis Health, omschrijft de stem als 'de sensor van het lichaam'. Zo heeft Vocalis Health een applicatie ontwikkeld waarmee mensen hun stem kunnen laten analyseren, en de progressie van bepaalde ziektes gemonitord wordt (Staff, 2020b). Tijdens de COVID-19 pandemie heeft het bedrijf een oproep gedaan om spraakopnames op te sturen, met als doel dragers van COVID-19 te signaleren (Staff, 2020a).

In samenwerking met een grote Amerikaanse verzekeringsmaatschappij analyseerde spraakbedrijf Canary Speech een bestand van miljoenen telefoongesprekken die gedurende vijftien jaar waren verzameld. Deze gesprekken werden geanalyseerd om de ziekte van Parkinson en de ziekte van Alzheimer te detecteren. Deze gegevens kunnen mogelijk gebruikt worden om verzekeringspremies aan te passen. Toen de algemeen directeur van de verzekeringsmaatschappij hiernaar gevraagd werd, gaf die aan dat zo'n applicatie gereguleerd zou moeten worden (Reynold, 2017).

Het bedrijf Voicesense geeft aan dat het op basis van spraakgegevens een volledig persoonlijkheidsprofiel kan opstellen (Chen, 2019). Het bedrijf biedt aan

gezondheidsklachten in te schatten, waaronder de kans dat mensen een aanvraag doen bij de verzekering, en de kans dat een patiënt terug naar huis kan. Ook analyseert Voicesense de kans dat klanten wanbetalingen doen en profileert het bedrijf werknemers, bijvoorbeeld ten aanzien van het risico op een burn-out. Ten slotte biedt Voicesense ook aan om te analyseren in hoeverre een klant een bepaald product trouw zal blijven.

In paragraaf 2.2.3 gaven we al aan dat stemanalyse nog in de kinderschoenen staat – de beloftes van de technologiebedrijven moeten daarom met een korrel zout genomen worden. Maar de ambities zijn groot.

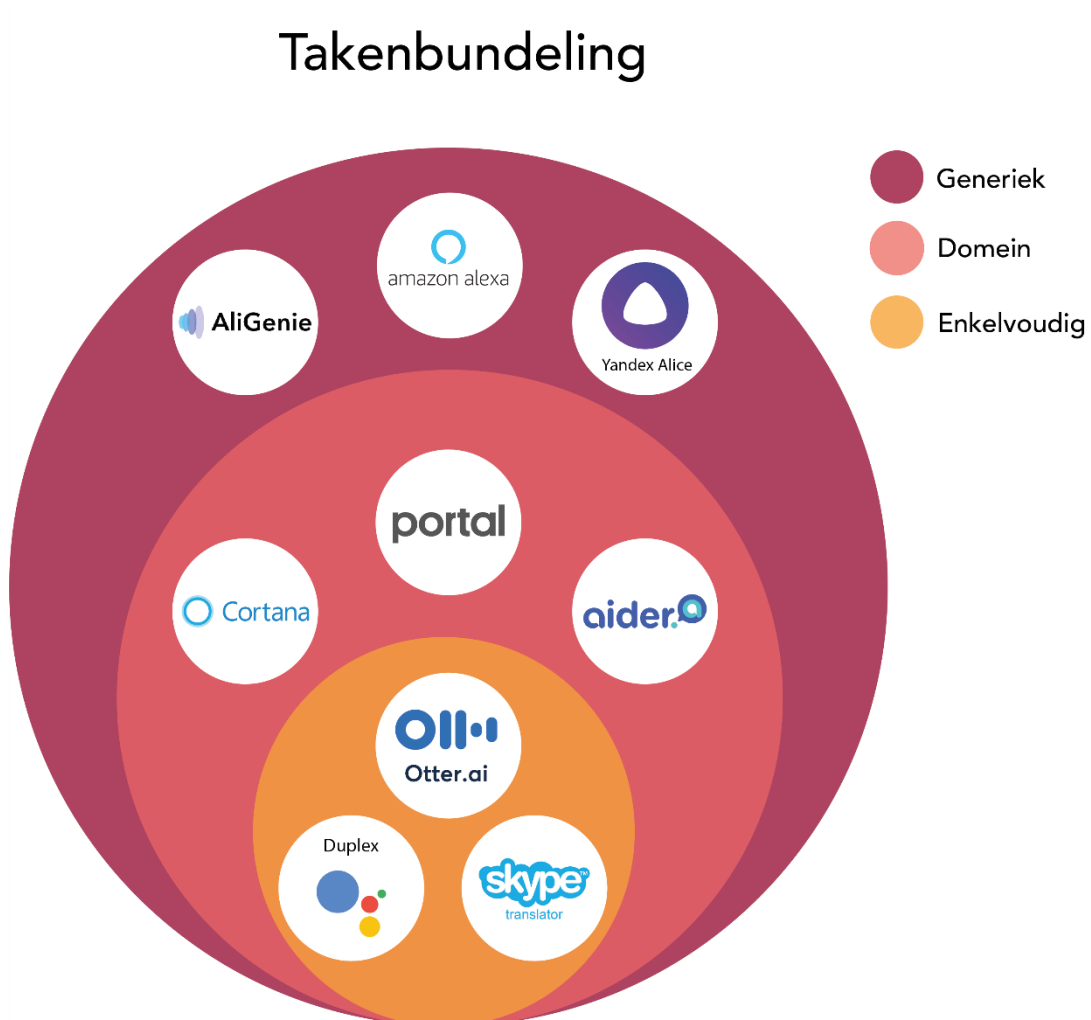
3.3 De markt voor spraaktechnologie

In de vorige paragrafen zien we een groeiende bedrijvigheid ontstaan op het gebied van spraaktechnologie. Deze paragraaf zoomt uit, en geeft een indruk van de markt voor spraaktechnologie. Wat zijn de ambities van deze bedrijven? Is de technologie vooral in handen van enkele grote partijen of zijn er alternatieven? Bij spraaktechnologie denken we vaak aan de spraakassistenten van de grote technologiebedrijven, Alexa van Amazon of Google Assistant. Maar er zijn meer aanbieders. Ze verschillen in de manier waarop ze **taken bundelen**, en in de manier waarop ze **applicaties ontwikkelen**.

Taakbundeling

We onderscheiden op het gebied van takenbundeling drie soorten spraaktoepassingen: generieke, domeinspecifieke en enkelvoudige (zie figuur 5).

Figuur 7 Takenbundeling



Inmiddels biedt vrijwel elke technologiereus een spraakassistent aan. Denk aan Amazon Alexa, Google Assistant en AliGenie, de assistent van het Chinese bedrijf Alibaba. Deze assistenten bundelen verschillende taken: ze kunnen muziek afspelen, de verwarming hoger zetten e.d.. Het zijn **generieke spraaktoepassingen**. Deze spraakassistenten streven ernaar om overal waar iemand is van dienst te zijn en ons door de digitale wereld te gidsen. Ook sommige kleinere bedrijven brengen generieke spraakassistenten op de markt, zoals Alice van het Russische bedrijf Yandex.

Er zijn ook assistenten die **domeinspecifiek** zijn. Zij richten zich op het uitvoeren van zoveel mogelijk taken binnen een bepaald domein. Een voorbeeld is de kantoorassistent Cortana van Microsoft, Portal van Facebook, die vooral focust op videobellen, Athena voor machines in de fabriekshal of de spraakdiensten van Aider.ai die vooral gericht zijn op kleinere bedrijven. Het voordeel is dat een

dergelijke assistent zich in een bepaald domein kan specialiseren, en zo tot betere dienstverlening kan komen. Toen bleek dat Microsoft met Cortana achterbleef bij de assistenten van de andere techreuzen, besloot het bedrijf tot deze gerichtere dienstverlening (Warren, 2019).

De grote technologiebedrijven richten zich op het verbeteren en uitbreiden van hun dienstverlening in specifieke domeinen. Daarmee gaan ze de concurrentie aan met bedrijven die actief zijn in deze specifieke domeinen. Amazon ontwikkelt bijvoorbeeld Alexa-diensten voor de zorg, als tegenhanger van Nuance.¹⁰ Google is in april 2020 gaan samenwerken met de medische spraakassistent Suki van het gelijknamige bedrijf.¹¹

Sommige spraaktoepassingen bieden een **enkelvoudige** toepassing: zij kunnen slechts één taak uitvoeren. Denk aan Google Duplex, die reserveringen maakt in restaurants. Vanwege de focus op deze ene taak, is Duplex daarin erg goed – zo komt de spraak van de software zeer menselijk over en kun je in een gesprekje met de software handig je voorkeuren aangeven. Maar andere taken dan reserveringen maken, doet Duplex niet. Naast Google Duplex kun je ook denken aan Skype Translator, en aan Otter.ai, dat bijvoorbeeld automatisch transcripties van toespraken genereert.

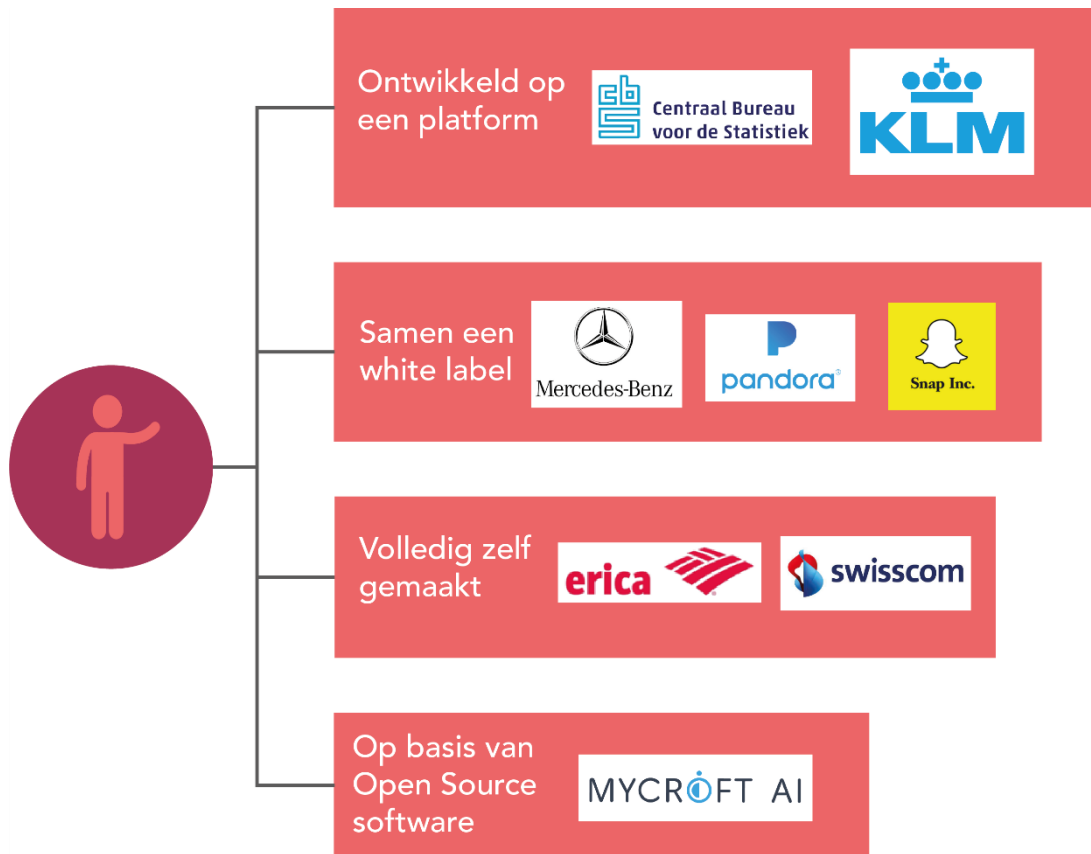
Productontwikkeling en aanbidding

Een ander verschil is de manier waarop de spraaktoepassing wordt ontwikkeld en aangeboden. Aanbieders kunnen spraaktoepassingen zelf maken, op een platform ontwikkelen, op *white label* software baseren of op *open source* software baseren (zie figuur 6).

10 <https://voicebot.ai/2020/06/11/nuance-connects-wolters-kluwer-clinical-database-to-dragon-medical-virtual-assistant/>

11 <https://voicebot.ai/2020/04/13/how-clinical-voice-assistant-startup-suki-lowers-insurance-claim-rejection-by-19/>

Figuur 8 Productontwikkeling en aanbieder



Diverse aanbieders ontwikkelen spraaktechnologie zelf, en bieden deze zelfstandig aan. Denk bijvoorbeeld aan Erica, de spraakassistent van de Bank of America en Swisscom, de assistent van het gelijknamige telecommunicatiebedrijf. Deze bedrijven hebben de software zelf ontwikkeld, en leveren deze rechtstreeks aan een klant.

Veel bedrijven en organisaties maken echter gebruik van de platformen van grote technologiebedrijven zoals Amazon en Google. Bedrijven en organisaties kunnen 'skills' en 'actions', toepassingen waarmee gebruikers met de bedrijven en organisaties kunnen interacteren, via het platform aanbieden. De 'informatiedialoog' van het CBS is hier een voorbeeld van, net als de spraaktoepassing van KLM. De bedrijven en organisaties maken in dit geval gebruik van een licentie van de spraaksoftware van grote technologiebedrijven.

Niet alleen de grote technologiereuzen bieden deze 'platformvariant'. Autofabrikanten Honda, Mercedes-Benz en Peugeot, maar ook andere bedrijven als Deutsche Telekom, Motorola, Snapchat en Pandora maken bijvoorbeeld voor hun spraaksoftware gebruik van het Houndify-platform van het bedrijf

Soundhound.¹² Dit platform is een *white label* product, wat betekent dat de software door Soundhound wordt gemaakt, maar door andere bedrijven met toevoeging van hun eigen merk en specificaties wordt verkocht. Dit geeft bedrijven meer controle over hun data en mogelijkheid om hun merk centraler te positioneren. Aan de andere kant profiteert het bedrijf niet van de rijke functionaliteit van een grote spraakassistent. Houndify biedt inmiddels meer dan veertien talen aan, waaronder Engels, Chinees, Spaans, Duits en Japans. Kleine bedrijven die gebruik willen maken van Houndify betalen via 'kredietpunten', waarbij elke vraag via Houndify een aantal kredietpunten kost. Een bedrijf kan er overigens ook voor kiezen om een deel van de functionaliteit uit te laten voeren door Houndify, en een deel door een andere assistent.

Ten slotte is er ook een *open source* alternatief, waarvan de broncode publiekelijk beschikbaar is gesteld en hoge privacystandaarden in acht worden genomen: het bedrijf Mycroft.¹³ Mycroft biedt een generieke spraakassistent die vrijelijk hergebruikt en aangepast kan worden, en zowel geïnstalleerd kan worden op Mycroft-apparaten als op andere hardware. Op dit moment is Mycroft wel vele malen kleiner dan de andere aanbieders van spraaktechnologie, en afhankelijk van donaties en crowdfunding. Het kan bij lange na niet de investeringen bijbenen van een bedrijf als Amazon. Bovendien beschikt het over minder trainingsdata, en verzamelt het vanwege de sterke privacyafspraken ook minder data (Newman, 2018). Het is dus de vraag of Mycroft qua dienstverlening, hardware en taalaanbod kan concurreren met de grote platformbedrijven. Het is nu in elk geval geen volwaardig alternatief.

De gids in de digitale wereld

Eén markttrend is duidelijk: de technologie-reuzen beconcurreren elkaar om het grootste spraakplatform aan te kunnen bieden. Die concurrentiestrijd is goed te begrijpen: hoe meer gebruikers en toepassingen het platform heeft, hoe aantrekkelijker het is – voor gebruikers, aanbieders én adverteerders. Dit wordt het netwerkeffect genoemd. De eigenaar van het platform kan met alle nieuwe data de spraakassistent nog verder verbeteren, en het platform nóg functioneler maken.

De ambitie van bedrijven als Google en Amazon is om zo'n breed inzetbaar platform te creëren, dat gebruikers het thuis, in de auto, op het werk, en waar ze verder ook zijn, willen inzetten. Het spraakplatform vormt een *seamless* omgeving, een gids die je als gebruiker altijd mee wilt nemen en die de digitale wereld ontsluit. Commentatoren omschrijven de strijd om dit grote, alomvattend spraakplatform als de *voice war* (Yoffie e.a., 2018 & Kinsella, 2019a). De Amerikaanse technologie-reuzen, zoals Google en Apple, lijken voorlopig in deze strijd voor te

¹² <https://www.houndify.com/>

¹³ www.Mycroft.ai

lopen op hun Chinese tegenhangers, zoals Baidu. De Chinese bedrijven zijn voornamelijk gericht op de Chinese markt, hetgeen alles te maken heeft met de Chinese taal (Harris, 2015).¹⁴

Om het breed inzetbare platform te bouwen, investeren de technologie-reuzen intensief in onderzoek en ontwikkeling (R&D). Zo heeft Amazon een Alexa-fonds opgericht dat 200 miljoen dollar investeert in kleinere bedrijven, en een Alexa-prijs in het leven geroepen die is gericht op het sponsoren van universitair onderzoek (Amazon, 2020). En Google heeft Google Assistants Investments opgericht, dat vooral partnerschappen aangaat in de zorg en op het gebied van toerisme en hotels (CBInsights, 2019). We merken hiervoor al op dat de grote technologiebedrijven ook in steeds meer domeinen actief worden, en de concurrentie aangaan met bedrijven die domeinspecifieke en eenvoudige spraakapplicaties maken.

Het is de vraag wat de concurrentiestrijd tussen de grote bedrijven op de langere termijn betekent voor de markt, en de brede groep van aanbieders die nu daarop actief is. Zullen kleinere aanbieders worden overgenomen door de grote technologiebedrijven of blijven er alternatieven voorhanden? In hoofdstuk vier gaan we nader in op voorwaarden voor een eerlijke, open en toegankelijke markt.

3.4 Conclusie

In dit hoofdstuk keken we in welke domeinen spraaktechnologie wordt toegepast. We zagen dat spraaksoftware al een breed aantal toepassingsgebieden kent. In de auto en in huis is spraaktechnologie het meest ontwikkeld. Maar ook in andere domeinen verkennen bedrijven, overheden en organisaties mogelijkheden, zoals telecom, bankieren, winkelen, de luchtvaart, de auto-industrie, de maakindustrie, de kantooromgeving, de zorg en het onderwijs. Hier is echter nog geen sprake van een brede toepassing. De toepassingen laten wel een duidelijke richting zien hoe de technologie tot nu toe wordt ingezet: spraaktechnologie wordt verkocht als onze gids, die tal van diensten ontsluit, gesprekken met ons voert, onze vragen beantwoordt en ons zelfs kan identificeren.

14 Deze markttrend is goed terug te zien in de cijfers. Wereldwijd hebben meer dan 2,5 miljard apparaten een ingebouwde spraakassistent. Het grootste deel hiervan bestaat uit de mobiele telefoons van Google en Apple, waarop de spraakassistenten van deze bedrijven, Google Assistent en Siri, zijn geïnstalleerd. Zo is de Google Assistent inmiddels geïnstalleerd op meer dan 1 miljard apparaten, en gebruiken 500 miljoen mensen deze assistent elke maand. Apple Siri is op meer dan 500 miljoen apparaten te vinden en kent 375 miljoen mensen die deze assistent maandelijks gebruiken. Maar ook andere spraakassistenten zijn te vinden op honderden miljoenen apparaten, zoals Microsoft Cortana, Baidu DuerOS, Amazon Alexa en Samsung Bixby (Canalys, 2020).

Het brede gebruik van spraaktechnologie gaat gepaard met een groeiende verzameling en analyse van spraakdata: de stem is een nieuwe databron, die nieuwe diensten en toepassingen mogelijk maakt. Gespreksverslagen, en vaak ook de audio-opnames van menselijke stemmen, vormen de basis van analyses die spraaktoepassingen mogelijk maken, (zoals het detecteren van ziekte), of verbeteren, (zoals het personaliseren op basis van iemands voorkeuren). In sommige gevallen worden spraakanalyses verkocht ten behoeve van het analyseren van gezondheid of het beoordelen van iemands werkprofiel. Zowel grote technologiebedrijven als kleinere start-ups zoeken intensief naar de stemkenmerken van ziektes en emoties. In het volgende hoofdstuk gaan we nader in op de maatschappelijke en ethische vragen die dit oproept.

De markt voor spraaktechnologie wordt gedomineerd door de grote technologiebedrijven, die forse investeringen doen en elkaar beconcurreren om het grootste spraakplatform te worden. Het nu al grote marktaandeel van deze bedrijven wordt verder vooruit geholpen door het netwerkeffect: het platform wordt aantrekkelijker voor toeleveranciers en gebruikers naarmate er meer toepassingen en gebruikers komen. Het doel is om met spraaktechnologie niet alleen een handige toegangspoort te creëren tot de digitale wereld, maar ook een gids die voortdurend bij de gebruiker blijft. Dit kan zijn tijdens het boodschappen doen of wanneer iemand iets opzoekt of naar muziek luistert. Een gebruiker hoeft het desbetreffende platform zo niet meer te verlaten.

Toch bestaan er ook andere platformomgevingen, gemaakt door *white label* en *open source* aanbieders, die bedrijven in staat stellen hun eigen toepassingen te bouwen en aan te bieden. De vraag is wat de toenemende concurrentiestrijd tussen de grote bedrijven op de langere termijn betekent voor de markt: (hoe) blijven er alternatieven voorhanden? Hier gaan we het volgende hoofdstuk op in.

4 Maatschappelijke en ethische aspecten van spraaktechnologie

Dit hoofdstuk biedt een overzicht van de maatschappelijke en ethische aspecten van spraaktechnologie. Het verkent welke betekenis spraaktechnologie heeft, en kan hebben, in de samenleving. Daarnaast brengt het de vragen die spraaktechnologie oproept in kaart. Eerst gaan we echter in op waarom onze spraak zo belangrijk voor ons is.

Het hoofdstuk is gebaseerd op het werk van ethici, media-onderzoekers en journalisten die de maatschappelijke en ethische aspecten van spraaktechnologie hebben geanalyseerd. Ook maken we gebruik van verschillende studies van het Rathenau Instituut, met name 'Opwaarderen' (2017) en 'Mensenrechten in het robottijdperk' (2017), die al aandacht vroegen voor de maatschappelijke en ethische aspecten van digitale technologie, waaronder robots en AI.

4.1 Het belang van spraak

Het vermogen om te spreken is op allerlei manieren essentieel voor de levens van mensen. Zo is onze spraak een **uiting van onze persoonlijkheid**: je kunt geduldig spreken, maar ook gehaast. Je kunt iets nuchter zeggen of juist vurig. Spraak is een van de belangrijkste uitingsvormen (Cox, 2019). Zo zegt een accent bijvoorbeeld iets over je achtergrond, net als de woordenschat die je gebruikt en de manier waarop je een zin formuleert. Je spraak vormt, samen met je uiterlijk, je visitekaartje.

Media-onderzoeker Sherry Turkle wijst ook op het belang van het gesprek tussen mensen: het ondersteunt empathie, vriendschap, liefde, leren en productiviteit (Turkle, 2017).

In de gesprekken die we met elkaar voeren **ontwikkelen we onszelf**. We wisselen kennis uit en bediscussiëren nieuwe ideeën. Een goed gesprek kan je op andere gedachten brengen of juist je overtuiging sterken. In onze gesprekken beschrijven we nieuwe plekken waar de ander nog niet is geweest, en geven we onze ervaringen door. Een gesprek kan ons leiden naar een wijdere wereld. En natuurlijk leren we ook over elkaar. We wisselen onze persoonlijkheden uit, en vormen in onze gesprekken een beeld van de ander.

Daarbij zijn we niet alleen uit op het verwerven van kennis. Want onze gesprekken zijn ook een dominante manier waarop we **ons tot elkaar verhouden**. In onze gesprekken komen we nader tot elkaar of raken we met elkaar in conflict, doen we elkaar na of respecteren we juist elkaars verschillen. In onze gesprekken staat dus een hoop op het spel. In een gesprek kan je een vriend maken of een vriend verliezen, opkomen voor je belangen of je belangen schaden.

Onze gesprekken zijn daarom, ten slotte, ook een dominante context voor **het ontwikkelen en toepassen van sociale normen en waarden** (Habermas, 1997). Het gesprek is de hoeksteen van een democratie, waarin we met elkaar de samenleving vormgeven. Deze normen zijn talrijk en alom aanwezig. Als je spreekt, of juist luistert, heb je je te gedragen. Je kunt niet zomaar door elkaar heen praten, je wordt geacht met aandacht naar elkaar te luisteren, en je hoeft geen geduld te hebben met sprekers die zich onnodig grof of kwetsend uiten. We hebben sociale normen over spreektijd, over zwijgen, over het geven van complimenten en het geven van kritiek. Al deze normen zijn voortdurend in ontwikkeling – denk bijvoorbeeld aan het gebruik van ‘u’. En ze passen zich aan uiteenlopende contexten aan – van een gezellig gesprek tussen vrienden tot een debat tussen politici.

Het mag duidelijk zijn: onze gesprekken zijn van belang voor onze sociale interactie en persoonlijke ontwikkeling. Door middel van hun spraak kunnen mensen zich ontwikkelen en uiten. **Weinig is meer intiem, en meer eigen, dan onze stem**. De woorden en zinnen die we spreken vormen de bouwstenen van onze persoonlijkheid. Bovendien geven we met spraak complexe en essentiële sociale processen vorm. Onze spraak is een kostbaar goed.

Juist omdat onze spraak zo belangrijk is, kan spraaktechnologie een grote maatschappelijke betekenis hebben. Spraaktechnologie mengt zich immers in onze gesprekken, en vormt in toenemende mate een onderdeel van onze sociale wereld. We praten met spraaktechnologie, en spraaktechnologie luistert naar ons. Dat roept vragen op. Niet alleen als de technologie naar behoren functioneert, maar ook als de spraaktechnologie fouten maakt.

Bedrijven zetten in op spraak als een interface tussen de mens en de digitale wereld. Maar hoe zou die interactie eruit moeten zien? Het is belangrijk om de ontwikkeling en implementatie van spraaktechnologie in goede banen te leiden. We signaleren daarbij vijf centrale ethische aspecten (zie figuur 7). Het is van belang om te zorgen dat spraaktechnologie:

- **sociale relaties en normen** versterkt, en niet vervaagt,
- **privacy** respecteert, en niet schendt,
- **autonomie** ondersteunt, en niet schaadt,

- **veilig** en **gezond** gebruikt kan worden, en
- leidt tot gebalanceerde **machtsverhoudingen** tussen bedrijven onderling, en bedrijven en burgers.

We zullen de maatschappelijke en ethische aspecten achtereenvolgens bespreken.

Figuur 9 Ethische aspecten van spraaktechnologie



4.2 Sociale relaties en normen

Spraaktechnologie kan steeds beter menselijke spraak nabootsen en deelnemen aan menselijke gesprekken. Dit betekent dat het verschil tussen computers en mensen vager wordt, waardoor onze relatie met computers verandert. Dat is geen neutrale zaak: spraaktechnologie kan onze sociale relaties en normen zowel schaden als verbeteren. Het is daarom belangrijk om de verandering in goede banen te leiden, en passende sociale normen te ontwikkelen. Vijf onderwerpen spelen daarbij een rol: het bevestigen of juist bestrijden van **vooroordelen**, het **behandelen van computers als mensen**, het **opvoeden en disciplineren** van

gebruikers, het **opnemen of buitensluiten** van gebruikers en het **overnemen van menselijk werk en contact**.

Vooroordelen

Het Rathenau Instituut constateerde al in het rapport '*Opwaarderen*' dat vooroordelen, zoals de opvatting dat echte vrouwen kinderen opvoeden, of het idee dat echte mannen heteroseksueel zijn, een hardnekkig probleem vormen bij de ontwikkeling van AI-technologie (Kool e.a., 2017). Als dergelijke opvattingen, onbedoeld of ongemerkt, in de technologie sluipen, kan dat leiden tot discriminatie of uitsluiting van bepaalde groepen (Barocas & Selbst, 2016). Ook bij spraaktechnologie spelen vooroordelen een rol (Howard & Borenstein, 2017). Zo is Unesco kritisch op het feit dat de namen van de meeste grote spraakassistenten, zoals Cortana en Siri, vrouwelijk zijn en ze standaard met een vrouwelijke stem spreken (UNESCO, 2019). Bovendien gedragen de spraakassistenten zich onderdanig en expressief – waarmee ze het vooroordeel voeden dat vrouwen zich altijd onderdanig en spontaan op moeten stellen.¹⁵

Het is dus wenselijk om vooroordelen in het ontwerp te ontmoedigen; het is echter niet direct duidelijk *hoe* dat het beste kan gebeuren. Genderneutrale stemmen zijn een optie, die niet klinken als een man of een vrouw (Simon, 2019). Of een elektronische stem, die we ook associëren met een computer. Beide keuzes ervaren we echter als minder prettig. Veel bedrijven geven daarom de gebruikers verschillende opties om een stemkeuze in te stellen, al wil dit niet zeggen dat vooroordelen daarmee verdwijnen. Een andere optie is om de stemmen van de assistent af te laten wisselen, om zo bewuster te interacteren met verschillende types. Het juiste antwoord zal niet gelijk gevonden zijn.

Computers behandelen als mensen

Het is bekend dat mensen in hun taalgebruik menselijke eigenschappen toeschrijven aan niet-menselijke objecten (Nass & Brave 2005; Ruane, e.a., 2019). Zo noemen we knuffels bijvoorbeeld 'lief', terwijl ze zich nooit lief zullen gedragen – ze zijn slechts lief in onze verbeelding. Dit wordt ook wel antropomorfisme genoemd: mensen projecteren menselijke eigenschappen op de wereld om hen heen. Dat doen mensen met auto's, computers, en ook met spraaktoepassingen.¹⁶ Sterker nog: ontwikkelaars willen spraaktechnologie zo ontwikkelen dat de software zo menselijk mogelijk overkomt.

15 Hoewel de software inmiddels is geüpdatet, beantwoordde Siri ooit de opmerking 'Siri, you're a bitch', met het antwoord 'I'd blush if I could'. De assistent antwoordt nu 'I don't know how to respond to that'.

16 Over deze grenserving zijn vele boeken geschreven en films gemaakt – denk bijvoorbeeld aan de film *Ex Machina*, waarin een programmeur de verleidingen van een levensechte robot moet weerstaan.

Dat kan de gebruikservaring verbeteren, zodat we met plezier en effectiviteit taken uitvoeren door middel van spraaktechnologie. Het toeschrijven van menselijke eigenschappen kan ons wellicht ook minder eenzaam maken of minder verdrietig. Maar het kan ook verwarrend en onprettig zijn als je computers en mensen niet meer uit elkaar houdt.¹⁷ Spraaktechnologie brengt grensvervaging dichterbij (Shank e.a., 2019). Het is niet denkbeeldig dat kwetsbare groepen zoals ouderen en kleine kinderen sommige huidige spraakassistenten, zoals bijvoorbeeld Google Duplex, niet meer van levende mensen kunnen onderscheiden en zich er zelfs emotioneel aan hechten.¹⁸ Ook een techniek als *voice cloning* kan verwarring en grensvervaging in de hand werken, evenals misbruik (dit bespreken we in paragraaf 4.5).

Als je menselijke gedrag geloofwaardig imiteert, komen menselijke reacties vanzelf naar boven. Zeker omdat spraaksystemen meestal geen robotlichamen hebben, waardoor we herkennen dat een computer aan het woord is.¹⁹ Spraaksystemen zijn, net als menselijke radiostemmen, stemmen die speakers en oortjes produceren. Dat roept ethische bezwaren op. Als mensen echt denken dat een computer een mens is, lijkt die leugen in te gaan tegen hun waardigheid – mensen voor de gek houden is niet respectvol. Dat roept vragen op: welke mate van toeschrijving van menselijkheid vinden we acceptabel of zelfs welkom? En op welk moment gaat de verwarring duidelijk te ver? Een minimale voorwaarde is dat computers altijd aangeven een computer te zijn.

Ongewenst taalgebruik en disciplineren

Spraaktechnologie kan onze stemmen opvangen, begrijpen, en erop reageren. Dat betekent ook dat spraaktoepassingen kunnen reageren op scheldwoorden, onbeleefdheid of op grammaticaal incorrecte zinnen. Dat roept verschillende vragen op. Ten eerste kan spraaktechnologie, net als bijvoorbeeld chatbots, verkeerde gewoontes aanleren, en bijvoorbeeld scheldwoorden gaan hanteren (Ruane e.a., 2019). Zo kregen gebruikers het voor elkaar om spraakassistent Siri bijvoorbeeld 'motherfucker' te laten zeggen (Leswing, 2018). Ontwikkelaars willen dat natuurlijk voorkomen, en zoeken naar manieren om spraaksystemen zo te trainen dat dergelijk misbruik onmogelijk is.

Een andere discussie gaat over het disciplineren van het taalgebruik van gebruikers door spraaksoftware. Een spraakassistent kan bijvoorbeeld de gebruiker vermanen

17 Deze discussie speelt al langer rondom de ontwikkeling en inzet van sociale robots, bijvoorbeeld in de zorg. In diverse rapporten over robots stelt het Rathenau Instituut de vraag tot welk niveau we de emotionele binding tussen mens en machines willen stimuleren en er gebruik van willen maken (Royakkers et al 2012; Van Est e.a., 2017).

18 En wat te denken van een baby die zijn of haar ouders niet aanspreekt met 'papa' of 'mama', maar met 'Alexa', zo speculeert columnist van Waarlo in de Volkskrant (Waarlo 2019)?

19 Er zijn wel spraaksystemen die belichaamd zijn, zoals Tinybot Tessa. Zie tinybots.nl.

als er gescholden wordt of beleefdheid juist belonen. Amazon heeft bijvoorbeeld in Echo smart speakers al software geïmplementeerd waarbij Alexa als je ‘please’ gebruikt, zegt ‘by the way, thanks for asking so nicely’ (Baig, 2019). Ook Google introduceerde ‘pretty please’ software. Mogelijk helpt de software bepaalde gewoontes tegen te gaan of helpt ze wellicht bij het opvoeden van kinderen. Maar het roept ook de vraag op of we dan een ethische grens overgaan: willen we dat robots ons disciplineren en opvoeden? Is die taak niet uitsluitend weggelegd voor mensen? En leren spraaksystemen ons ook slechte gewoontes aan, zoals elkaar met enige frustratie commando’s geven?

Bij sommige menselijke interacties kun je de vraag stellen of robots een rol moeten spelen – zoals bij een oplopende echtelijke ruzie of bij een ouder die een woedend kind probeert te kalmeren. Er is al software die een seintje geeft bij te hoge decibellen, een (mogelijk) teken van ruzie (Kool e.a., 2014). Samenvattend betekent het dat het respecteren van menselijke waardigheid een belangrijk uitgangspunt moet zijn bij de ontwikkeling spraaktechnologie (zie hiervoor ook de studie ‘Opwaarderen’ van het Rathenau Instituut, Kool e.a., 2017).

Gebruikers in- of uitsluiten

Spraaktechnologie kan om verschillende redenen bepaalde groepen beter bedienen dan anderen, onder andere vanwege de taal die de software beheerst. Spraaktoepassingen herkennen en spreken de ene taal beter dan de andere, afhankelijk van hoe ze getraind zijn. Omdat bepaalde markten interessanter zijn dan andere, en er van bepaalde talen veel meer data beschikbaar zijn, kunnen spraaksystemen doorgaans beter Engels en Spaans spreken dan Nederlands en Pools. En binnen bepaalde talen beheerst het systeem de manier van spreken van de meerderheid beter dan het accent van de minderheid. De software werkt immers vooral goed als mens en omgeving meewerken: er moet duidelijk gearticuleerd worden, met bekende woorden, in een rustige omgeving en in een taal die het systeem goed beheerst (zie hoofdstuk twee).

Dit kan uitsluiting in de hand werken, en druk zetten op minderheden om zich te conformeren aan het taalgebruik van de meerderheid. Deze dynamiek is al zichtbaar in het wereldwijde taallandschap, waar al veel talen en dialecten zijn verdwenen (Dujardin, 2017). Ook andere vormen van uitsluiting zijn denkbaar, zoals spraaksystemen die vrouwen slechter verstaan dan mannen, kinderen slechter verstaan dan volwassenen of die voor sommige groepen gebruikers simpelweg te ingewikkeld zijn (Tatman, 2017).

Ontwikkelaars zijn natuurlijk niet uit op uitsluiting – maar het is voor hen aantrekkelijk om zich op dominante taalgebieden te focussen, en bias (ofwel een ingebouwd vooroordeel of vooroordelen) treedt in het trainen van AI-systemen al

snel op. Bovendien is het lang niet altijd eenvoudig deze vooroordelen op te sporen of te verhelpen (Barocas & Selbst, 2016). Bijvoorbeeld omdat het vanwege privacyoverwegingen lastig is om goede datasets van kinderspraak op te bouwen. Naast meer bewustzijn is daarom een gerichte inzet op het bevorderen van inclusie nodig.

Overnemen van menselijk werk en contact

In het voorgaande hoofdstuk zagen we dat spraaktoepassingen, naast ondersteuning in het persoonlijke domein, ook werk met en contact met mensen kan overnemen, bijvoorbeeld in grotendeels geautomatiseerde callcenters. Spraaktechnologie raakt zo aan de maatschappelijke en politieke discussie over de toekomst van werk, waarbij het behouden en creëren van kwalitatief goed werk een centrale vraag is. In plaats van alleen te kijken hoe mensenwerk kan worden vervangen door technologie, vraagt dat van bedrijven een bredere blik om te zien hoe mensen en technologie elkaar zo mogelijk kunnen aanvullen en verrijken.

Het inzetten van spraaktechnologie, bijvoorbeeld bij klantcontactcentra of in de zorg, betekent ook dat wij in ons dagelijks leven te maken zullen krijgen met spraakrobots. Hoewel dat soms op bepaalde momenten een uitkomst zal zijn, kan het ons ook schaden. Worden we wel goed geholpen, en kan technologie wel goed voor ons 'zorgen'? Is *zorgen voor* niet bij uitstek iets dat mensen elkaar geven of moeten geven? Het Rathenau Instituut pleitte eerder al voor een recht op menselijk contact, waarmee mensen kunnen kiezen of ze door een (ro)bot te woord willen worden gestaan of door een mens (Van Est e.a., 2017).

4.3 Respect voor privacy

Spraaktechnologie kan op steeds meer plekken onze gids zijn, maar zo ook met ons meeluisteren, ons aanspreken en data over ons verzamelen. In de Verenigde Staten, waar spraaktechnologie het hardst groeit, zijn er zelfs gebruikers die meer dan tien slimme speakers in huis hebben geïnstalleerd, zodat ze vanuit elke kamer met spraak hun digitale toepassingen kunnen bedienen (Kinsella, 2019b).

Dat roept vragen op over de bescherming van onze privacy. Juist op de plekken waar we ons van oudsher onbespied wanen, en onbevange ons zelf kunnen zijn, zoals ons huis en auto, luistert spraaktechnologie nu mee. Spraaktechnologie raakt op meerdere manieren aan ons privéleven: door het **verzamelen** en **gebruiken** van persoonlijke gegevens, door de **inmenging** van spraaktechnologie in ons privéleven en door het effect van spraaktechnologie op onze **zelfexpressie**.

Dataverzameling

Allereerst roept spraaktechnologie de vraag op of onze data correct worden verzameld. Als iemand ruzie maakt met zijn of haar geliefde dan is die ruzie privé. Maar wat als er een slimme speaker in de slaapkamer staat? Wie mag er meeluisteren? En op welk moment? En welke informatie mag worden opgehaald?

Deze vraag draait grotendeels om het zogeheten ‘wake word’. Het idee is namelijk dat gebruikers van spraakassistenten zelf de technologie activeren, door bijvoorbeeld ‘hey Siri’ of ‘hey Google’ te zeggen. Als de gebruiker het niet wil, luistert er dus niemand mee.

Maar dit systeem is niet foutloos. Soms komt het voor dat een gebruiker per ongeluk de technologie activeert, doordat je iets zegt dat lijkt op “hey Siri” (Hern, 2019; Cox, 2020). Dat kan meerdere keren per dag gebeuren. Bovendien neemt het systeem een paar seconden voordat je iets zegt al op, omdat hij anders het ‘wake word’ niet kan horen, en luistert hij een paar seconden langer mee dan strikt noodzakelijk, om niks te missen (Shulevitz, 2018). Hoezeer ontwikkelaars het ook proberen te voorkomen – het komt wel degelijk voor dat spraaktechnologie meeluistert, terwijl dit niet de bedoeling is.

Een van de wettelijke uitgangspunten is de eis om zo min mogelijk data te verzamelen, en verzamelde data zo snel mogelijk te vernietigen (het principe van dataminimalisatie). Dit verkleint het risico op misbruik. Bij spraak zou dat onder andere betekenen dat niet het audiogeluid, maar alleen de omgezette tekst verzameld wordt, zodat een commando correct kan worden uitgevoerd. Maar ontwikkelaars willen controleren of de audio goed omgezet wordt in tekst – en dat betekent dat hun werknemers luisteren naar een selectie van verzamelde audiobestanden (Day, e.a., 2019). Deze controles kunnen al snel privacyschendingen opleveren. Zo rapporteerde een klokkenluider die werkte voor Apple dat hij luisterde naar geluidsopnames van mensen die seks hadden, meeluisterde bij drugdeals en ook medische details hoorde (Hern, 2019).

Datagebruik

Privacy vereist niet alleen dat onze informatie correct wordt verzameld, maar ook correct wordt gebruikt. De Algemene Verordening Gegevensbescherming (AVG) formuleert hiervoor de belangrijke uitgangspunten. Naast dataminimalisatie zijn dat bijvoorbeeld toestemming, informatie geven en doelbinding: dataverwerking moet gebonden zijn aan een vooraf vastgesteld doel.²⁰ Dat klinkt helder, maar alle drie leveren ze in de praktijk vragen op – bijvoorbeeld als opsporings- en

20 AVG artikel 5.

veiligheidsdiensten beslag willen leggen op de data. Zo overwegen de Duitse autoriteiten om data op te eisen om de opsporing van criminaliteit te ondersteunen, en in de Verenigde Staten is dit al mogelijk (Reuters, 2019; Statt, 2018; Reilly, 2016). We zagen in hoofdstuk drie dat ook Interpol een stemdatabase heeft aangelegd. Wanneer staat een democratische samenleving dit soort verzoeken toe, en wanneer niet? Hier wordt nu nog nauwelijks debat over gevoerd.

Gebruikers beseffen niet altijd hoe hun gegevens door aanbieders, en andere partijen, worden geanalyseerd. Welke informatie en profielen worden vastgelegd? En op welke manier worden deze gebruikt? Momenteel is dat vooral om te leren hoe de persoonlijke interactie met een systeem kan worden verbeterd. Maar de stem fungeert steeds meer als nieuwe databron. Ons stemgeluid is heel waardevol. Het zal aantrekkelijk zijn voor ontwikkelaars om de doelen van dataverwerking op te rekken, gezien de opkomende handel in stemanalyses (zie hoofdstuk drie). Allerlei partijen, van kredietbeoordelaars, en adverteerders tot verzekeraars, zullen geïnteresseerd zijn in deze analyses en profielen. Dit vraagt om een actieve toezichthouder, om de principes die al wettelijk vastgelegd zijn zo goed mogelijk te handhaven.

Daar hoort ook een fundamentele discussie bij over hoe mensen meer grip krijgen op profilering. Hoewel de AVG bescherming biedt, kan er nog steeds profilering plaatsvinden. Het Rathenau Instituut riep daarom in het rapport 'Mensenrechten in het robottijdperk' op tot een recht om niet gemeten, geanalyseerd en beïnvloed te worden (Van Est e.a. 2017).

Ook aanbieders kunnen zich nog meer inspannen om het gebruik van stemgeluid toch tot het minimum te beperken. Hierbij zijn verschillende oplossingsrichtingen te bedenken. Je kunt bijvoorbeeld stemacteurs inhuren om een spraaksysteem te trainen of gebruik maken van *federated learning*, waarbij de data niet centraal opgeslagen worden, maar op de apparaten van gebruikers blijven staan.

Inmenging in ons privé- en familieleven

Het recht op respect voor ons privé- en familieleven is vastgelegd in de Grondwet internationale verdragen, waaronder het Europese Verdrag voor de Rechten van de Mens (De Grondwet Art. 10, EVRM Art. 8). Spraaktechnologie kan – naast het openbaar maken of lekken van persoonlijke gegevens – zich ook mengen in ons privéleven. Bijvoorbeeld als de spraakassistent, door een foutief opgevangen *wake word*, inbreekt in een gevoelig gesprek of ruzie. Of als een gezinslid ongevraagd de spraakassistent erbij haalt. Ons privéleven is gevuld met sociale interacties die veranderen als er een computer bij betrokken wordt. Dit kan ook positieve aspecten hebben. Bij spraaktechnologie onttrekken gezinsleden zich bijvoorbeeld niet aan

een gesprek door naar de tv of mobiel te kijken – het gesprek wordt gezamenlijk gevoerd.

Zelfexpressie

Een belangrijk onderdeel van privacy is de mogelijkheid om in vrijheid je eigen identiteit te uiten (Koops, 2019; Amnesty International, 2019). Als mensen weten dat ze in de gaten gehouden worden, kunnen er *chilling effects* optreden – mensen voelen zich dan niet vrij om zichzelf te uiten. Dit zou ook voor kunnen komen in de context van spraaktechnologie. Als er overal spraaktechnologie aanwezig is, bij mensen thuis maar ook in de publieke ruimte, kunnen *chilling effects* dus ook overal optreden. Stel je eens voor dat je niet vrijuit durft te spreken in je auto, of zelfs in je woonkamer – dat is een dystopisch beeld. Toch is onbekend welke *chilling effects* precies door spraak optreden, en hoe die zich verhouden tot andere digitale technologie, zoals camera's.

4.4 Autonomie

Spraaktechnologie belooft gebruikers te helpen om handelingen uit te voeren en beslissingen te nemen. De naam assistent zegt het al: een handige hulp waarmee we kunnen winkelen, reserveren, gesprekken voeren en vragen kunnen beantwoorden. Gemakkelijk en efficiënt worden we door de (digitale) wereld geloosd. Tenminste, dat is de belofte. Maar een gids heeft ook een andere kant: een gids laat ook dingen niet zien en maakt beslissingen voor ons. Spraaktechnologie kan zo onze vaardigheden en autonomie beïnvloeden.

Vaardigheden

Een terugkerende discussie bij de inzet van technologie is de manier waarop het onze vermogens beïnvloedt. We leren er bepaalde taken bij, en kunnen andere verliezen of minder goed uitvoeren – *deskilling* (Danaher, 2018). Als we de wereld met spraak kunnen bedienen, kan dat uitkomst bieden voor laaggeletterden of mensen met een visuele beperking (Arcon e.a., 2017). Het kan ook betekenen dat als we meer gaan spreken, we minder schrijven en daar ook minder goed in worden. We noemden aan het begin van dit hoofdstuk al mediaonderzoeker Sherry Turkle. Zij vindt in haar jarenlange onderzoek aanwijzingen dat de interactie via schermen – in plaats van rechtstreeks met elkaar – onze empathie vermindert (Turkle, 2017). Turkle betoogt dat we het gesprek moeten *reclaimen*, ofwel opnieuw moeten opeisen, en meer met elkaar moeten praten. Het is nog te vroeg om te weten hoe de opkomst van spraaktechnologie, en het voeren van een gesprek, onze relaties zal beïnvloeden. Die veranderingen zijn niet per se slecht. De vraag is welke waarde de vaardigheid die we verliezen precies heeft, en hoe we ervoor zorgen dat we in de nieuwe situatie beter af zijn (Danaher, 2018).

Sturen van informatie

Informatie opzoeken verloopt via spraak anders dan via een scherm. Op een scherm is doorgaans meer informatie te zien: meerdere zoekresultaten, waarbij diverse meetindicatoren en achterliggende bronnen al zichtbaar zijn. Ook bij een Wikipedia-pagina of socialmediabericht is aanvullende informatie te zien over de betrouwbaarheid of de afzender. Bij spraak zijn deze mogelijkheden er doorgaans minder. Ontwikkelaars proberen spraakassistenten vaak een enkel antwoord te laten geven, bijvoorbeeld op de vraag hoeveel katholieken er in Nederland zijn. Dan noemt de assistent een cijfer op, zonder te specificeren of het gaat om katholieken die ook naar de kerk gaan of om katholieken die ingeschreven staan bij de kerk, en van welke bron de informatie afkomstig is. Hoewel het handig is om snel een antwoord te krijgen, waarbij de gebruiker snel leert, is de vraag wat voor antwoord dit precies is. Er kan een versimpelde versie van de werkelijkheid worden gegeven, waarbij de gebruiker uiteindelijk weinig leert (Vlahos, 2019). Bovendien is de vraag hoe het antwoord tot stand komt, en wie dat bepaalt. Is dat het eerste zoekresultaat? Antwoorden worden via spraak op deze manier meer gestuurd, en gecureerd, dan onze interactie via schermen.

Een andere ontwerpkeuze is om de spraakassistent een dialoog te laten voeren. Zo heeft het Centraal Bureau voor de Statistiek (CBS) met Google technologie een spraaktoepassing gemaakt die een 'informatiedialoog' kan voeren met de gebruiker. In deze dialoog is meer nuance mogelijk, wat noodzakelijk is gezien de complexe dataset die het CBS beheert. Ook bij het zoeken naar medische informatie is het van belang om specificaties en nuanceringen in de dialoog naar voren te laten komen. Dit is alleen wel complexer om te maken, en het kan meer kennis van de gebruiker vergen.

Sturen van beslissingen

Met de informatie die spraaktechnologie ons geeft, kunnen spraakassistenten ook helpen bij het nemen van beslissingen. Bijvoorbeeld welke route te volgen of bij het samenstellen van het boodschappenlijstje. Idealiter ondersteunt de technologie het menselijke denkproces: iemand streeft naar een bepaald doel, en de spraakassistent maakt het mogelijk dat doel te bereiken. Iemand wil naar Den Haag, en het navigatiesysteem kiest de snelste route uit.

Maar soms is de verhouding tussen wat iemand wil en wat spraaktechnologie aanbiedt complexer, en kan beïnvloeding plaatsvinden (Stucke & Ezrachi, 2016). Bijvoorbeeld als de spraakassistent al denkt te weten wat we willen en suggesties doet. Denk bijvoorbeeld aan een assistent die helpt bij het maken van boodschappenlijstjes, en een bepaald menu suggereert. Die suggesties hebben invloed op het denkproces en de beslissing van een gebruiker – en dat kan hem of

haar minder autonoom maken. We merkten in de vorige paragraaf al op dat een respectvolle assistent gebruikers niet onbewust of oneigenlijk beïnvloedt.

Sturen van gebruikerservaring

Spraaktechnologie is niet alleen maar een middel om iets anders te bereiken, maar levert ook een ervaring op. Dit is het meest zichtbaar bij spraakassistenten die speciaal ontwikkeld worden om als gezelschap te dienen. De klant betaalt voor een gebruikservaring. De digitale wereld kent inmiddels vele aantrekkelijke producten en diensten, die burgers intensief gebruiken. Ook hier hoeft niets mis mee te zijn. Als een digitale dienst leuk is en plezier aan ons leven toevoegt, is dat positief. Maar aantrekkingskracht kan omslaan in verslaving. Dat is zeker bij spraaktechnologie denkbaar. Het kan *te* leuk of *te* handig zijn om van spraakassistenten gebruik te maken – ze staan immers 24/7 voor een gebruiker klaar (Kretzschmar e.a., 2019). We kunnen meer met spraaktechnologie interacteren dan goed is voor onze gezondheid.

Deze verslaving is treffend verbeeld in de film *Her*, over een man die verliefd wordt op zijn spraakassistent, en in series zoals *Black Mirror*. Dit zijn natuurlijk sciencefiction-scenario's, maar er zijn al concrete ontwikkelingen die zorgen baren. De chatbot Replika is bijvoorbeeld puur gericht op persoonlijk contact, en kan een gesprek over schoonheid voeren (Harris, 2019). Zo'n chatbot kan al verslavend werken.²¹ Ook bij de introductie van de sociale robot Jibo bleek dat gebruikers een emotionele band met een pratende robot kunnen opbouwen, die verslaving in de hand kan werken (Van Camp, 2019). Ten slotte worden er ook spraakgames ontwikkeld, en kan spraaktechnologie ook al verslavende toepassingen nog aantrekkelijker maken, zoals virtual reality games (Conner, 2020). Het is dus niet ondenkbaar dat we net zo verknocht raken aan onze spraaktechnologie als aan onze mobiele telefoons.

Sturen van nieuws

Deze ontwerpkeuzes zijn ook van belang voor de nieuwsvoorziening, en de discussie over desinformatie. De recente studie van het Rathenau Instituut 'Digitalisering van het nieuws', naar het tegengaan van desinformatie, benadrukt het belang van een pluriform media-aanbod én mediaconsumptie (Van Keulen e.a., 2018). In Nederland raadplegen de meeste mensen meerdere nieuwsbronnen per dag. Spraakassistenten bieden een nieuw kanaal om nieuws tot je te nemen, en het nieuws geregeld 'af te spelen'. Gebruikers kunnen de bron zelf kiezen, bijvoorbeeld NOS of nu.nl. Een gebruiker krijgt dan bijvoorbeeld elk uur in een bericht van drie minuten de hoogtepunten van het nieuws te horen. Net als bij het zoeken op het web via een scherm, is er minder ruimte beschikbaar. En dus worden ook hier

21 Zie voor een voorbeeldgesprek: <https://www.youtube.com/watch?v=eOqeSC28EhU>.

keuzes gemaakt over wat de gebruiker in die drie minuten te horen krijgt. Vooralsnog wordt dit nieuws niet gepersonaliseerd en maken gebruikers ook gebruik van andere nieuwsbronnen. Op dit moment heeft de 'skill' nieuws afspelen dus niet direct invloed op de verspreiding van desinformatie.

Maar op andere vlakken bestaan er wel degelijk zorgen. Spraaktechnologie biedt nieuwe mogelijkheden om valse berichten te creëren, bijvoorbeeld via 'voice cloning'. In juni 2020 voegde Twitter als eerste grote socialemediaplatform een spraakfunctie toe: gebruikers kunnen een audiofragment opnemen en op hun tijdlijn plaatsen. Net als vragen rondom *deep fakes*, is de vraag hoe platformen deze berichten op echtheid kunnen controleren. Met betrekking tot het verspreiden van desinformatie zijn veel maatregelen op sociale media erop gericht om aanvullende informatie bij berichten te geven. Bij spraak zal hier over het algemeen minder ruimte voor zijn – hoe gaan aanbieders en beleidsmakers met deze beperking om?

4.5 Veilig en gezond gebruik

Sinds de komst van computers is cyberveiligheid een punt van aandacht. Elke computer kan gehackt en gecorrumpeerd worden – en er zullen helaas altijd partijen geneigd zijn om hier misbruik van te maken. Bovendien kan elke computer fouten maken en een groot of klein ongeluk veroorzaken. Daarnaast kan spraaktechnologie gezondheidsschade opleveren. We bespreken hier daarom drie elementen: de bescherming tegen **kwaadwillende actoren**, het **betrouwbaar maken** van spraaktoepassingen en het **voorkomen van gehoorschade**.

Kwaadwillende actoren en beveiliging

Kwaadwillende actoren, zoals cybercriminelen en vijandige inlichtingendiensten, kunnen op verschillende manieren misbruik maken van spraaktechnologie. Allereerst kan spraaktechnologie gehackt worden, waardoor iemand toegang krijgt tot bepaalde spraakdata of mee kan luisteren via een spraaktoepassing. Een crimineel kan dan bijvoorbeeld een wachtwoord afluisteren en geld stelen of burgers chanteren met gevoelige informatie. Bovendien kan met gestolen spraakdata iemands stem nagemaakt worden (*voice cloning*), waarmee weer ander misbruik mogelijk wordt, zoals identiteitsfraude. Bedenk dat bedrijven inzetten op biometrische stem-IDs, juist om de veiligheid van hun diensten te verbeteren. Maar als deze stem-IDs nagemaakt worden, staat daarmee de veiligheid van die nieuwe identificatiesystemen onder druk.

Een ander gevaar schuilt in zogeheten *adversarial examples* (Alzantot e.a., 2018). Dit zijn data waarmee *deep learning*-systemen om de tuin geleid kunnen worden, zoals een plaatje van een kat dat zo bewerkt is, dat de computer het toch als een

hond categoriseert. Er wordt op dit moment onderzoek gedaan naar de mogelijkheid om ook spraaktechnologie om de tuin te leiden, bijvoorbeeld door bepaalde achtergrondgeluiden toe te voegen. Onderzoekers claimen dat ze spraaktechnologie met geluiden kunnen besturen, die mensen met hun eigen oren niet waar kunnen nemen (Smith, 2018).

Deze risico's vragen om maatregelen. Zo wordt er onderzoek gedaan naar trainingsmethodes die AI-systemen beschermen tegen gekloonde stemmen en *adversarial examples* (Wu, e.a., 2020). Maar het is vrijwel onmogelijk om digitale apparaten helemaal veilig te maken. Zelfs in de meer betrouwbare systemen, zoals het besturingssysteem van IOS van Apple, worden met regelmaat kwetsbaarheden ontdekt, zoals de publicatie van het Rathenau Instituut 'Cyberspace zonder conflict' liet zien (Hamer e.a., 2019). Toch valt er winst te behalen. Van sommige apparaten, zoals de huishoudelijke apparaten die een groot deel uitmaken van het Internet of Things, is bekend dat de cyberveiligheid te wensen overlaat. Als spraaktechnologie op meer van deze apparaten geïnstalleerd wordt, kan het veiligheidsprobleem snel groter worden.

Betrouwbare toepassingen

Onveilige situaties kunnen ook per ongeluk ontstaan, als de technologie fouten maakt of ons te veel afleidt. Dat is niet erg als het een klein incident betreft, bijvoorbeeld als vertaalsoftware een lidwoord vergeet te vertalen. Maar als spraaktechnologie belangrijke functies vervult in een auto of medische informatie doorgeeft, kan een technische fout of een moment van onoplettendheid ernstige gevolgen hebben (Strayer, 2015). We schreven al dat spraaktechnologie snel beter wordt, zeker wat het herkennen van spraak betreft. Maar foutmarges van 5% zijn onacceptabel bij kritische toepassingen – en vaak is de foutmarge veel hoger.

De inzet van spraaktechnologie in specifieke domeinen vraagt daarom om een verdere daling van de foutmarges en om aandacht voor het effect van spraaktechnologie op onze concentratie – voordat deze in de praktijk wordt ingezet. Aan de andere proberen bedrijven met spraaktechnologie processen juist veiliger te maken – bijvoorbeeld doordat spraakassistenten bij bepaalde processen een checklist aan vragen voorleggen, zodat een zorgmedewerker, of iemand die een industriële machine bestuurt, niets vergeet (Softengi, 2020).

Voorkomen van gehoorschade

Spraaktoepassingen kunnen ook leiden tot gehoorschade. Dit speelt met name als de spraaktechnologie in een oortje of koptelefoon zit of via een oortje of koptelefoon wordt beluisterd. Bij langdurig gebruik en te hoge volumes kunnen oortjes en koptelefoons namelijk gehoorschade veroorzaken (RTL Nieuws, 2019). Deze

schade is niet kinderachtig: oorartsen waarschuwen dat met name jongeren gehoorschade oplopen waar ze een leven lang last van houden. Gehoorschade is meestal niet te genezen. Het is niet denkbeeldig dat de opkomst van spraaktechnologie dit groeiende probleem verder verergert. De filosofie van spraakontwikkelaars, en in het bijzonder spraakassistenten, luidt immers dat de technologie voortdurend in contact staat met de gebruiker en overal mee naartoe gaat. Dan kan het voor de gebruiker aantrekkelijk zijn om de oortjes veel in te hebben, en wordt de kans op gehoorschade groter.

4.6 Macht van technologiebedrijven

In hoofdstuk drie zagen we dat de grote technologiebedrijven de markt voor spraaktechnologie domineren. Vrijwel elke technologieus heeft zijn eigen assistent ontwikkeld, die op een groeiend aantal apparaten te gebruiken is. De techbedrijven zetten voornamelijk in op assistenten die als een multifunctionele gids veel taken, ook in specifieke domeinen zoals de zorg, goed kunnen uitvoeren. Ze investeren fors in onderzoek en kopen start-ups op (CBinsights, 2019). Er bestaat daarom een reële kans dat de markt voor spraaktechnologie in handen komt van slechts een aantal bedrijven. De technologiebedrijven versterken zo hun toch al machtige positie, die maatschappelijk en politiek steeds meer ter discussie komt te staan.

De bedrijven die spraaktechnologie verkopen kunnen op verschillende manieren macht uitoefenen over de burgers die hun producten en diensten kopen. Zo proberen bedrijven hun gebruikers te verleiden om binnen hun commerciële ecosysteem allerlei diensten te kopen – denk aan de spraakassistent Alexa die gekoppeld is aan de online winkel van Amazon. De techbedrijven streven ernaar zoveel mogelijk interacties (zoeken, winkelen, reizen, betalen) binnen het eigen domein te laten plaatsvinden. Consumenten kunnen natuurlijk besluiten om toch met een ander bedrijf in zee te gaan. Maar door diensten en producten te koppelen, oefenen bedrijven wel degelijk invloed uit op burgers – er ontstaat een lock-in-effect (Van Dijck e.a., 2016). Op die manier wordt het lastiger voor gebruikers om over te stappen naar een andere ontwikkelaar.

Beleidsmakers in Nederland, Europa en elders proberen op verschillende manieren de groeiende marktmacht van de technologieuzen tegen te gaan. Deels is dat gericht op actievere handhaving van bestaande wet- en regelgeving, en deels zoeken beleidsmakers naar het aanpassen van diverse juridische kaders, waaronder mededinging, consumentenrecht, en privacyrecht.

Maar er is meer nodig. We schreven al dat de technologiereuzen door hun machtige positie veel gegevens van gebruikers verwerken – in een groeiend aantal domeinen. Kunnen kleinere partijen daar nog wel mee concurreren? En hoe zorgen we ervoor dat talen, accenten en doelgroepen die de grote bedrijven niet interessant vinden toch worden meegenomen? Dit lijkt niet alleen te vragen om regelgeving, maar ook om actieve overheidsinvesteringen in spraaktechnologie die spraaksystemen inclusiever maakt.

4.7 Conclusie

Dit hoofdstuk gaf een overzicht van de maatschappelijke en ethische aspecten die verbonden zijn met de opkomst van spraaktechnologie. Op basis van dit overzicht concluderen we dat **de spraak van mensen een essentieel goed is**. Door middel van hun spraak ontwikkelen en uiten mensen zich, verhouden ze zich tot elkaar en worden belangrijke sociale normen in stand gehouden of ontwikkeld. Het gebruik van spraaktechnologie zal de interactie tussen mensen onderling, tussen mensen en apparaten, en tussen bedrijven en instellingen veranderen. Spraaktechnologie vormt een gids die mensen door de digitale wereld leidt. Dit hoofdstuk besprak daar diverse voorbeelden van, zowel positief en negatief, en liet zien welke ethische en maatschappelijke vragen deze gids oproept. Niet alle vragen zijn gemakkelijk of gelijk te beantwoorden met 'goed' of 'fout'. We weten nog niet precies hoe menselijke relaties precies gaan veranderen, hoe spraaktechnologie zich mengt in het privéleven, en hoe vaardigheden en beslissingen precies worden beïnvloed.

Maar er zijn wel een aantal minimumvoorwaarden waar de inzet van spraaktechnologie aan moet voldoen. Duidelijk is dat ontwikkelaars, bedrijven en andere instellingen, de stemmen, gesprekken en sociale processen van mensen in acht moeten nemen. Vijf ethische, maatschappelijke en juridische aspecten spelen daarbij met name een rol: het respecteren van **sociale normen**, het beschermen van **privé- en familielevens**, het bevorderen van de **autonomie**, het zorgen voor **een veilig en gezond** gebruik en het creëren van **eerlijke en toegankelijke markt** voor spraaktechnologie.

5 Conclusie: tijd voor een goed gesprek

Spraaktechnologie heeft de afgelopen jaren een grote ontwikkeling doorgemaakt. Ten opzichte van tien jaar geleden kunnen computers menselijke spraak veel beter herkennen, interpreteren en spreken – hoewel de gebruiker hier vaak nog wel een handje bij moet helpen. Sommige spraakcomputers zijn soms zelfs moeilijk van mensen te onderscheiden. Inmiddels bevindt spraaktechnologie zich overal om ons heen: in onze huiskamers, in onze auto's, in onze laptops en in onze telefoons.

Onze spraak vormt een essentieel onderdeel van wie we zijn als mens, en van onze sociale relaties. Ook bevatten onze gesprekken zeer gevoelige informatie – over onze identiteit, over de gesprekken die we voeren, en zelfs over onze gezondheid en ons humeur. Onze spraak verdient dus bescherming. Spraaktechnologie moet onze stemmen versterken, en niet misbruiken of discrimineren. Spraaktechnologie moet onze relaties – met elkaar én met computers – verrijken, en niet verstoren. En spraaktechnologie moet onze economie eerlijker maken, en niet de dominantie van grote technologiegiganten vergroten.

Wat is er voor nodig om dit te bereiken? In dit slothoofdstuk werpen we de blik vooruit en doen we voorstellen. Maar eerst vatten we de belangrijkste bevindingen van de voorgaande hoofdstukken nog even samen.

5.1 De stand van zaken

5.1.1 Spraaktechnologie wordt volwassen

Hoofdstuk twee liet zien dat spraaktechnologie bestaat uit drie centrale processen: het **herkennen** van spraak, het **interpreteren** van spraak, en het spreken zelf, dat **spraaksynthese** wordt genoemd.

Spraakherkenning is vaak behoorlijk goed, met, in ideale omstandigheden, foutmarges van rond de 5%. Voor veel toepassingen is deze mate van spraakherkenning voldoende om een nuttige dienst te verlenen, bijvoorbeeld als het gaat om het op afstand bedienen van muziek of het transcriberen van een

interview. Toch is deze foutmarge significant – er zijn genoeg toepassingen te bedenken, zoals in de zorg of zware industrie, waarbij een dergelijke foutmarge niet acceptabel is. Bovendien neemt de foutmarge toe in een rumoerige omgeving.

Op het gebied van **spraakinterpretatie** is de voortgang minder eenduidig. Bij het uitvoeren van taken is vaak veel hulp van de omgeving en de gebruiker nodig: hij of zij moet precies de juiste commando's geven en vragen op de juiste manier formuleren en beantwoorden. Ook informatie op websites moet op een bepaalde manier toegankelijk worden gemaakt voor spraaktechnologie. Hoewel de belofte was dat computers onze taal gingen leren, moeten mensen zich nog steeds aanpassen aan de spraaktechnologie, wil deze goed geïnterpreteerd worden.

De **spraaksynthese** is, net als spraakherkenning, veel beter geworden. Kortgezegd kunnen spraaksystemen zich duidelijk verstaanbaar maken. Het doel van de ontwikkelaars ligt nu hoger: spraaksynthese moet zo goed zijn, dat mensen niet meer horen dat ze met een computer spreken. Bij verreweg de meeste toepassingen is dit niet het geval, en is te horen dat er spraaktechnologie wordt gebruikt. Maar de ontwikkelingen gaan hard – sommige specifieke spraaksystemen, zoals Google Duplex, komen heel dichtbij het produceren van menselijke spraak, inclusief 'uhhs' en 'mmhmmh's'.

Spraaktoepassingen beheersen steeds meer talen, met name grote taalgebieden zoals Engels, Spaans en Chinees. Op dit moment biedt alleen Google een assistent aan die de Nederlandse taal beheerst. De vooruitgang op het gebied van spraaktechnologie in het afgelopen decennium is met name te danken aan de opkomst van *deep learning*-systemen, die met behulp van de verbeterde rekenkracht van computers, en de groeiende verzameling van spraakdata, intensief getraind kunnen worden. Om verdere voortgang te boeken, en computers bijvoorbeeld nieuwe talen aan te leren, zijn meer data nodig: meer opnames van menselijke stemmen en gespreksverslagen. Een andere stap is spraaktoepassingen ook in lastigere omgevingen taken te laten uitvoeren, en zo invoelend mogelijk te laten communiceren. Daarvoor willen ontwikkelaars spraak combineren met andere gegevens, bijvoorbeeld gezichtsuitdrukkingen of lipbewegingen.

5.1.2 Spraaktechnologie wordt steeds meer gebruikt

In hoofdstuk drie lieten we zien dat diverse toepassingen van spraaktechnologie al breed in de samenleving wordt toegepast. Daarnaast zagen we technologieaanbieders en bedrijven volop experimenteren met spraak. We bespraken spraaktechnologie die apparaten aanstuurt en spraaktechnologie die

diensten ondersteunt of overneemt. De eerste categorie omvat spraaktechnologie in huis, in de auto, in de industrie en in wearables. De tweede categorie omvat spraaktechnologie op het gebied van reizen en toerisme, winkelen, in de zorg, het onderwijs, callcenters, de nieuwsvoorziening, taalwerk, identificatie en verificatie, online zoeken en advertenties. In al deze domeinen wordt spraaktechnologie verkocht als een behulpzame assistent, die tal van diensten ontsluit, gesprekken met ons voert, onze vragen beantwoordt en ons zelfs kan identificeren.

Al deze toepassingen verzamelen data, door middel van gespreksverslagen én audio-opnames. De stem fungeert zo als nieuwe databron. De data worden door ontwikkelaars gebruikt om spraaksystemen te personaliseren, en vormen de basis van analyses op het gebied van emotieherkenning en het diagnosticeren van ziekten. Die analyses zijn vaak niet wetenschappelijk bewezen, maar diverse bedrijven verwachten veel van de toekomstige mogelijkheden van audio-opnames.

Tenslotte gaven we aan dat spraaktechnologie steeds vaker onze gids is in de digitale wereld. De doelstelling van verschillende technologiegiganten zoals Google en Amazon is om dit portaal in handen te krijgen. Dat doen ze door een breed platform van spraaktoepassingen te creëren, en die te koppelen aan een spraakassistent, zoals Alexa en Google Assistant, die een veelheid aan taken kan uitvoeren. Ze kopen daarbij kleine start-ups op, en werken samen met partijen die veel kennis hebben van een specifiek domein, bijvoorbeeld de zorg. Hoewel er ook kleine partijen actief zijn op de markt voor spraaktechnologie, is de vraag hoe deze partijen zich staande houden in het 'gevecht om spraak' en ten opzichte van de groeiende machtspositie van de techgiganten.

5.1.3 Maatschappelijke en ethische aspecten van spraaktechnologie

Hoofdstuk vier zette uiteen welke maatschappelijke en ethische aspecten verbonden zijn met de opkomst van spraaktechnologie. Het maakte duidelijk dat spraak een van de belangrijkste manieren is waarmee mensen zich uiten, zich ontwikkelen en zich tot elkaar verhouden. Het is dus belangrijk dat de stemmen, gesprekken en sociale processen van mensen gerespecteerd worden. We bespraken vijf centrale aspecten:

1. **Sociale relaties en normen.** Spraaktechnologie mengt zich in het sociale leven van mensen. Dit heeft verschillende elementen: het aanmoedigen of juist afleren van vooroordelen, de grensvervaging die ontstaat als computers behandeld worden als mensen, de sociale disciplineren van gebruikers en het opnemen en buitensluiten van gebruikers. Hier ontstaan vragen over welke relatie we tussen mens en computers wenselijk achten: willen we, en

moeten we, altijd weten dat we tegen computers spreken in plaats van een mens? Wanneer vinden we het problematisch als gebruikers hun spraakassistent beschouwen als hun beste vriend? Welke taken, bijvoorbeeld opvoeden of zorg verlenen, willen we overlaten aan of laten uitvoeren door computers? Hoe zorgen we dat spraaktechnologie bestaande sociale normen, bijvoorbeeld ten aanzien van gelijke behandeling en disciplinerende, respecteert?

2. **Autonomie.** Spraaktechnologie helpt om taken uit te voeren, beslissingen te nemen en gebruikers een aantrekkelijke ervaring te geven. Maar er bestaan zorgen over het verliezen van vaardigheden, het beïnvloeden en sturen van gebruikers als zij naar informatie zoeken op het web, winkelen of nieuws tot zich nemen. Spraaktechnologie bevat in vergelijking met schermen doorgaans minder mogelijkheden voor nuance en minder informatie over de bron. Wie stuurt en bepaalt welk antwoord de spraakassistent geeft? Hoe zorgen we dat de gebruiker niet oneigenlijk wordt beïnvloed? Tot slot noemden we de mogelijk verslavende werking van spraaktechnologie.
3. **Privacy.** Spraakdata zijn zeer gevoelige data: juist in gesprekken thuis, in de auto en op het werk geven mensen zichzelf bloot. Bovendien wordt technologie ontwikkeld om uit onze stemmen aanvullende informatie te halen – naast wat we zeggen, dus ook *hoe* we dat zeggen. Dat kan ook worden gebruikt om ziekten op te sporen. Spraak als nieuwe databron vraagt daarom om extra aandacht van ontwikkelaars en toezichthouders, om te zorgen dat ons privé- en familielevens gerespecteerd blijven.
4. **Veilig en gezond gebruik.** Spraaktechnologie kan de veiligheid van mensen op verschillende manieren onder druk zetten. Spraakdata kunnen gestolen en misbruikt worden, bijvoorbeeld door een stem te klonen. En ondanks de verbeteringen, is spraaktechnologie niet perfect – in de aansturing van toepassingen kunnen ongelukken gebeuren. Voordat spraaktechnologie in kritieke toepassingen in de zorg, defensie of de maakindustrie wordt ingezet, zal de betrouwbaarheid van de technologie buiten kijf moeten staan. Ten slotte kunnen spraaktoepassingen leiden tot gehoorschade, als oortjes te veel en te luid worden gebruikt.
5. **Macht van technologiebedrijven.** Spraaktechnologie heeft een nieuwe en snelgroeïende markt gecreëerd. De bekende techgiganten domineren deze markt, en breiden hun bestaande machtspositie verder uit via spraak. De grote technologiebedrijven spelen dus als de belangrijkste ontwikkelaars ook een centrale rol in het respecteren van sociale normen, ons privéleven, en het creëren van veilige toepassingen die burgers niet schaden, verwarren, oneigenlijk beïnvloeden of verslaven.

5.2 Spraaktechnologie vraagt om maatschappelijke en politieke actie

Deze verkenning bouwt voort op een uitgebreide reeks studies die het Rathenau Instituut uitvoerde naar ontwikkelingen in de digitale samenleving. Zo wees het essay 'Intieme technologie' (2014) erop dat technologie steeds meer over ons weet, meer tussen ons in komt te staan en meer op ons gaat lijken. Het rapport 'Mensenrechten in het robottijdperk' (2017) wees op de impact van digitale technologie op mensenrechten als privacy, veiligheid en non-discriminatie. In de overzichtsstudie 'Opwaarderen' (2017) bracht het Rathenau Instituut de volle breedte van de ethische en maatschappelijke vraagstukken van de digitale samenleving in kaart, en werd gewezen op de verantwoordelijkheden van overheden, zoals gemeentes, provincies, de rijksoverheid, toezichthouders en uitvoeringsorganisaties, maar ook het bedrijfsleven en burgers zelf, om deze vraagstukken op te pakken.

Veel van de in deze studies aangekaarte kwesties en aanbevelingen zijn nog steeds actueel. Spraaktechnologie voegt een nieuwe dimensie toe aan de brede opdracht om digitale technologie in goede banen te leiden, en laat zien dat de overheid en het bedrijfsleven wederom aan zet zijn. Onze spraak staat immers op het spel. Onze stemmen en gesprekken vormen een essentieel onderdeel van wie we zijn als mens, en van de relaties die we met anderen aangaan. Met spraaktechnologie krijgen we er op allerlei momenten – thuis, in de auto, op het werk en bij het winkelen – een gesprekspartner bij, en dat zal onze spraak en onze relaties – met elkaar én met computers – niet onberoerd laten. Bovendien creëert spraaktechnologie een nieuwe bron van data, die zeer gevoelige gegevens bevatten.

Deze ontwikkelingen raken ons als individuele burgers, maar hebben ook gevolgen voor de samenleving als geheel. Spraaktechnologie beïnvloedt namelijk niet alleen de individuele omgang met computers, maar ook de omgangsvormen die we samen ontwikkelen. Het verandert niet alleen de manier waarop individuen kennis vergaren, maar ook de kennisbasis van het publieke debat. En het heeft niet alleen impact op de relatie tussen klanten en bedrijven, maar ook op de platformeconomie als geheel.

Het is aan de politiek om kaders te bieden waarbinnen spraaktechnologie ontwikkeld kan worden. Het is aan het bedrijfsleven om maatschappelijke verantwoorde technologie te maken en te beheren. En het is aan burgers om met elkaar het gesprek aan te gaan over de rol van spraaktechnologie in hun leven. Het rapport presenteert daarom een reeks aanbevelingen voor de overheid en het bedrijfsleven (zie tabel 1).

Tabel 1 Aanbevelingen

Aanbevelingen	Overheid	Bedrijfsleven
1. Zorg voor effectieve privacybescherming	Voer vergunningsplicht voor stemanalyse in. Monitor gebruik spraaktechnologie door opsporingsdiensten.	Geef privacybeginselen krachtig vorm
2. Bevorder inclusieve spraaktechnologie	Investeer in Nederlandse spraakdatabase. Spreek bedrijven aan op hun verantwoordelijkheden.	Waak voor stereotypering in stemgebruik en bevorder herkenning divers taalgebruik.
3. Creëer een eerlijke markt	Scherp het mededingingsrecht aan. Bied mogelijkheden voor alternatieve aanbieders.	Zet de rechten van consumenten centraal.
	Overheid en bedrijfsleven	
4. Bescherm menselijke waardigheid	Start ethische dialoog over spraaktechnologie en kom tot afspraken.	
5. Zorg dat spraaktechnologie betrouwbaar is	Pak desinformatie aan. Verlaag de foutmarges van spraaksystemen.	
6. Investeer in technologisch burgerschap	Leid burgers op om verantwoord om te gaan met spraaktechnologie, en stimuleer onderzoek.	

1. Zorg voor effectieve privacybescherming

Overheid: voer een vergunningsplicht voor stemanalyse in, en monitor het gebruik van spraaktechnologie door opsporingsdiensten.

Spraaktechnologie maakt het mogelijk om gevoelige gegevens van mensen te verzamelen en te gebruiken om mensen te beïnvloeden. Daaronder vallen ook zogeheten bijzondere persoonsgegevens, zoals biometrische gegevens waarmee iemand geïdentificeerd kan worden of gegevens over iemands gezondheid. De verzameling van stemdata levert daarom risico's op voor de mens en diens fundamentele rechten. Foutieve verwerking in de zorg kan leiden tot misdiagnose, de diefstal van een stem-ID kan leiden tot materiële schade, en de toenemende kans op identificatie kan leiden tot *chilling* effecten: iemand durft niet meer vrijuit te spreken uit angst dat diens stem geanalyseerd en/of geïdentificeerd wordt.

Niet voor niets benoemt de Europese Commissie (2020) het gebruik van biometrische toepassingen voor identificatie (op afstand) een voorbeeld van

toepassingen met 'hoog risico'. Een ethisch verantwoorde omgang met stemdata behoort tot de kern van 'ethische AI', een belangrijk uitgangspunt van zowel het huidige kabinetsbeleid als het beleid van de Europese Commissie ten aanzien van AI. Vanwege deze gevoeligheden gelden voor de verwerking van bijzondere persoonsgegevens in de huidige wetgeving (de Algemene Verordening Gegevensbescherming, AVG) al strengere waarborgen.²² De verwerking van bijzondere persoonsgegevens is in principe verboden, tenzij er sprake is van een van tien uitzonderingsgronden, en de verwerking voldoet aan de principes van proportionaliteit en subsidiariteit. Het is aan de toezichthouder om onrechtmatige dataverwerking op te sporen – en dat is gezien de talrijke toepassingen van bijzondere persoonsgegevens een zeer veeleisende taak.

Om het huidige verbod op de verzameling en verwerking van bijzondere persoonsgegevens effectiever te handhaven pleit het Rathenau Instituut voor het **invoeren van een vergunningsplicht voor biometrische analyse ten behoeve van verificatie en identificatie, waaronder stemanalyse** (Rathenau Instituut, 2020). Alle partijen die biometrische gegevens zouden willen verzamelen en verwerken met het oog op verificatie en identificatie, dienen dan vooraf een vergunning aan te vragen bij de Autoriteit Persoonsgegevens. De toezichthouder doet dat al met betrekking tot bepaalde verwerkingen van strafrechtelijke gegevens. De Uitvoeringswet AVG biedt mogelijkheden om een dergelijke vergunningsplicht ook voor biometrische gegevens in te voeren. Ook is het goed als de overheid strategieën ontwikkelt om sentimentanalyse en gezondheidsanalyse op basis van stemdata te reguleren.

Naast het invoeren van een vergunningsplicht is het ook van belang dat de overheid **de inzet van spraaktechnologie door opsporingsdiensten monitort**. Zo heeft Interpol al een aanzienlijke database met stemprofielen aangelegd, deels op basis van tijdens de opsporing onderschepte audio, en deels op basis van socialmediaberichten. Het gebruik van spraaktechnologie door opsporingsdiensten is wereldwijd in opkomst. Het is zaak om goed te kijken naar de dataverzameling van opsporingsdiensten. Is het bijvoorbeeld wenselijk als de politie stemdata van sociale media 'schraapt'? Veel burgers zullen hier bij het uploaden van hun stem geen rekening mee houden. Ook is het belangrijk om na te gaan of het toezichtssysteem voldoet: wordt effectief gecontroleerd of een opsporingsdienst zich aan de eigen regels houdt? Met de opkomst van spraaktechnologie verdienen deze vragen aandacht van de politiek.

²² Zie voor een uitleg van het wettelijk kader door de AP, de Handleiding Algemene verordening gegevensbescherming: autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/handleiding_avg.pdf.

Bedrijfsleven: geef privacybeginselen krachtig vorm.

Op dit moment worden er, met name door de AVG, allerlei eisen gesteld aan het datagebruik en de algoritmes van de ontwikkelaars van spraaktechnologie. Het is belangrijk dat het bedrijfsleven die regels niet minimaal volgt, maar zo krachtig mogelijk beginselen zoals transparantie, toestemming en dataminimalisatie vormgeeft, en daar zijn innovatieve vermogen ook op gericht. Zo is het goed als het bedrijfsleven investeert in technieken als *federated learning*, waarbij AI-systemen ontwikkeld worden zonder dat de data een apparaat verlaten, en de werking van spraakassistenten inzichtelijker maken voor de consument en de 'wake word-functie' verbeteren, die essentieel is voor het geven van toestemming.

2. Bevorder inclusieve spraaktechnologie**Overheid: investeer in een Nederlandse spraakdatabase en spreek het bedrijfsleven aan op zijn verantwoordelijkheden.**

Spraaktechnologie kan uitsluitende effecten hebben, omdat het stemmen van uiteenlopende groepen, zoals mensen die een dialect spreken, mensen op leeftijd of kleine kinderen, minder goed herkent en daardoor slechter werkt. Dat is zorgelijk, aangezien spraaktechnologie juist een kans biedt om de digitale wereld te ontsluiten voor minder mediawijze mensen, zoals ouderen. Bovendien zijn spraaksystemen die de taal van bepaalde groepen beter herkennen dan die van andere discriminerend – terwijl iedereen die zich in Nederland bevindt recht heeft op een gelijke behandeling. Ten slotte kan het te weinig bedienen van specifieke groepen een verschromelend effect hebben voor het Nederlands taalgebied, omdat de technologie mensen dwingt om een bepaald soort Nederlands te gaan spreken.

De overheid kan op verschillende manieren inclusie bevorderen en discriminatie tegengaan. Allereerst kan de overheid investeren in een rijke en diverse **Nederlandse taaldatabase**, die publiekelijk beschikbaar wordt gesteld. Op basis van die database kunnen vervolgens allerlei organisaties, bedrijven en de overheid zelf inclusieve spraaktechnologie ontwikkelen. Een dergelijke database is ook van belang voor het creëren van eerlijke en toegankelijke marktvoorwaarden (zie aanbeveling 3).

Daarnaast kan de overheid het bedrijfsleven aanspreken op zijn verantwoordelijkheden. Zo is het zaak dat ontwikkelaars van spraaktechnologie, net als andere aanbieders van diensten en producten, plannen ontwikkelen en beleid voeren om discriminatie tegen te gaan – het is aan de overheid om te monitoren of deze verantwoordelijkheid voldoende serieus wordt genomen, en daarover in gesprek te gaan met ontwikkelaars.

Bedrijfsleven: waak voor stereotypering.

Spraakassistenten gaan steeds meer met gebruikers in gesprek, en krijgen menselijke profielen. Het is daarom belangrijk dat ontwikkelaars ervoor waken dat deze **profielen geen stereotypes aanmoedigen**, zoals het stereotype dat vrouwen zich onderdanig en opgewekt moeten opstellen. Het zou goed zijn als er zo veel mogelijk diversiteit bestaat in het aanbod van stemmen en profielen, en als gebruikers hier ook keuzes in kunnen maken. Gezien het toenemende gemak waarmee een stem gesynthetiseerd kan worden, ligt dit binnen de technische mogelijkheden.

3. Creëer een eerlijke markt**Overheid: scherp het mededingingsrecht aan, en biedt mogelijkheden voor alternatieve aanbieders.**

Er zijn op dit moment grote zorgen over de machtspositie van enkele grote technologiebedrijven binnen de platformeconomie. Spraaktechnologie biedt deze grote bedrijven de mogelijkheid om hun machtspositie nog verder uit te breiden, wat kan leiden tot monopolievorming en de benadeling van andere aanbieders die de markt op komen, zoals *white label*-platformen. Het is daarom zaak dat de overheid stappen neemt om ervoor te zorgen dat de markt voor spraaktechnologie eerlijk is.

Dit betekent allereerst dat het **mededingingsrecht aangescherpt moet worden**. In Europees verband zoekt de overheid al naar aanscherping van het mededingingsrecht om een eerlijkere en meer toegankelijke markt te creëren. Ten aanzien van spraak betekent dit dat beleidsmakers extra alert moeten zijn op hoe spraak de bestaande dominantie van grote platformbedrijven verder kan vergroten. Daarnaast is het van belang dat de overheid en door de overheid gefinancierde organisaties als **launching customer** bij het aankopen van spraaktechnologie niet alleen kijkt naar de grote technologiebedrijven, maar ook naar *white label*-aanbieders. In de toekomst kan *open source* technologie wellicht uitkomst bieden. Om deze twee opties te laten slagen, is het van belang dat de overheid investeert in een **publieke spraakdatabase** met de Nederlandse taal, waar alle partijen gebruik van kunnen maken— zie aanbeveling 2.

Bedrijfsleven: zet de rechten van consumenten centraal.

Een eerlijke markt heeft niet alleen betrekking op verhoudingen tussen bedrijven maar ook de relatie tussen bedrijven en consumenten. Het is van belang dat bedrijven rechten van consumenten, zoals verhaal kunnen halen en informatie kunnen opvragen over de werking van producten en diensten, ruimhartig en slagvaardig vormgeven. Als bedrijven hun verantwoordelijkheid niet nemen, is het aan overheden, en met name toezichthouders, om consumentenrechten te

beschermen. Onlangs zijn daarvoor in Europa ook meer mogelijkheden gekomen.²³ Maar het belangrijkste is dat bedrijven in eerste instantie zelf de rechten van consumenten een centrale plek geven.

4. Bescherm menselijke waardigheid

Overheid en bedrijfsleven: start een ethische dialoog over de inzet van spraaktechnologie en kom tot afspraken.

Gezien de maatschappelijke en ethische kwesties die spelen rondom spraaktechnologie, en de rol van de overheid als hoeder van grond- en mensenrechten, is het van belang dat de overheid een dialoog met het bedrijfsleven opzet over de verantwoorde ontwikkeling en inzet van spraaksystemen. Twee onderwerpen vragen daarbij in het bijzonder om aandacht.

Allereerst kan spraaktechnologie **het recht op menselijk contact** onder druk zetten. Met de opkomst van spraaktechnologie wordt het mogelijk om gesproken dienstverlening niet door mensen, maar door computers uit te laten voeren. Bij de overheid, en bij allerlei bedrijven, kan dit in de toekomst bijvoorbeeld betekenen dat mensen eerst met een computer te maken krijgen als zij telefonisch contact zoeken. Het Rathenau Instituut pleitte eerder al voor het recht op menselijk contact: mensen dienen de keuze te krijgen of zij geholpen willen worden door een mens (Van Est e.a., 2017). De overheid kan dit uitgangspunt in ieder geval hanteren voor haar eigen dienstverlening, en ze kan andere partijen stimuleren om dit ook te doen.

Daarnaast is het van belang dat spraaktechnologie mensen **niet in verwarring brengt**. Dit betekent ten minste dat het voor gebruikers altijd duidelijk is of een computer spreekt of een mens. De overheid en het bedrijfsleven kunnen hier **afspraken over maken**. Ook is het van belang om een dialoog te voeren over de mate waarin computers empathisch zouden moeten communiceren met mensen, aangezien het projecteren van menselijke eigenschappen, en zelfs verslaving in de hand kan werken.

5. Zorg dat spraaktechnologie betrouwbaar is

Overheid en bedrijfsleven: pak desinformatie aan en verlaag de foutmarges van spraaksystemen.

²³ https://ec.europa.eu/commission/presscorner/detail/nl/IP_18_3041

Spraaktechnologie kan de samenleving veel bieden, maar moet wel betrouwbaar zijn. Om daarvoor te zorgen zijn verschillende stappen nodig. Allereerst is het van belang om **slagvaardig op te treden tegen desinformatie** en *voice cloning*. Overheden, en met name de Europese Commissie, hebben de afgelopen jaren met diverse maatregelen ingezet op het bestrijden van desinformatie. Een daarvan is om platformen nadrukkelijker wijzen op hun verantwoordelijkheden en de noodzaak zichzelf te reguleren. Zo is een Code of Practice on Disinformation afgesproken die door grote platformbedrijven is ondertekend. Toch is er aanhoudende kritiek op de platformbedrijven, en wordt hen verweten te weinig effectief op te treden. Daarom zou de overheid minder vrijblijvende maatregelen moeten overwegen, en de code zo moeten aanpassen dat de naleving met dwangmiddelen gehandhaafd kan worden. Ook kan geïnvesteerd worden in een systeem van digitale handtekeningen, zodat de authenticiteit van audio-opnames inzichtelijk wordt. Het is uiteindelijk aan de platformen zelf om desinformatie te detecteren. Het is zaak dat de bedrijven daartoe blijven inzetten op innovatieve technieken om desinformatie te onthullen en te verwijderen.

Ten tweede is het belangrijk dat het bedrijfsleven de **foutmarges van spraaksystemen verlaagt**. Deze marges zijn nog steeds significant: rond de 5% in goede omstandigheden en hoger in slechte omstandigheden. Dit betekent dat spraaksystemen die allerlei taken uitvoeren, makkelijk fouten kunnen maken en schade kunnen veroorzaken. Hoewel ontwikkelaars werken om deze marges te verkleinen, moet de overheid monitoren in hoeverre de hoge foutmarges veiligheidsrisico's opleveren en de ontwikkeling van veiligheidsstandaarden ondersteunen – zeker als spraaktechnologie meer gebruikt wordt bij kritische toepassingen zoals een medische operatie. Bovendien proberen kwaadwillende partijen om de spraakherkenning om de tuin te leiden, bijvoorbeeld door middel van *adversarial examples*. Het is daarom zaak dat overheden en bedrijven investeren in nieuwe technologie waarmee dit soort misbruik kan worden voorkomen.

6. Investeer in technologisch burgerschap

Overheid en bedrijfsleven: Leid burgers op om verantwoord om te gaan met spraaktechnologie, en stimuleer onderzoek.

Spraaktechnologie kan in sommige situaties een uitkomst zijn – zeker als mensen moeite hebben om met behulp van lezen en schrijven digitale toepassingen aan te sturen. Maar een verantwoord en effectief gebruik van spraaktechnologie vereist ook kennis en vaardigheden, bijvoorbeeld bij het zoeken naar informatie en het instellen van routines. De overheid en het bedrijfsleven kunnen verschillende stappen nemen om het technologisch burgerschap te ondersteunen.

Ten eerste is het zaak dat de overheid in samenwerking met bedrijven en maatschappelijke organisaties burgers **leert omgaan met spraaktechnologie**, en investeert in opleidingen en trainingen op het gebied van mediawijsheid. Burgers moeten zo leren om spraaktechnologie in te zetten en kennis opdoen over risico's zoals *voice cloning*. Ten tweede kunnen overheden en bedrijven structureel **onderzoek stimuleren** naar de invloed van spraaktechnologie op onze mentale en fysieke gezondheid.

Burgers: laat je stem horen, en zet spraaktechnologie op de agenda van het publieke debat.

Voor burgers zelf is er ook een belangrijke rol weggelegd. Dit rapport bracht tal van belangrijke maatschappelijke en ethische kwesties in kaart die publiek debat verdienen, zoals de inclusiviteit van spraaktechnologie, de relatie tussen mensen en sprekende computers, de impact van spraaktechnologie op onze sociale normen en de inzet van stemanalyse door bedrijven en overheden. Maar het is uiteindelijk ook aan burgers om, vanuit hun ervaringen met spraaktechnologie, deze thema's op de publieke agenda te zetten. Dit rapport eindigt daarom met een oproep: deel je ervaringen, en spreek je uit! Onze spraak is een kwetsbaar en betekenisvol goed – en een discussie meer dan waard.

5.3 Slotwoord

Dit onderzoek maakt onderdeel uit van een groep studies naar digitale technologieën die onze relatie met de digitale wereld veranderen. Naast spraaktechnologie is in de studie 'Verantwoord virtueel' ook onderzoek gedaan naar Virtual Reality (VR), en analyseerde de studie 'Nep echt - Verrijk de wereld met Augmented Reality' de technologie Augmented Reality. Deze technologieën staan niet los van elkaar, maar kunnen worden gecombineerd – bijvoorbeeld als je een AR-bril aanstuurt met je stem. Daarom is het van belang om in gesprek te gaan over de optelsom van deze technologieën: de synthese van spraaktechnologie, VR en AR. Dat gesprek zal het Rathenau Instituut organiseren en aanjagen. Want juist als we achterhalen hoe burgers deze technologieën gebruiken, en welk effect deze technologieën op ons hebben, kan de maatschappij het gebruik van digitale technologie die ons leven verrijkt bevorderen, en het gebruik van digitale technologie die ons leven verarmt tegengaan.

Literatuur

3PlayMedia. (2019). *State of Automatic Speech Recognition. An Annual Report*.
<https://go.3playmedia.com/rs-2019-asr>

Abdulsatar, A. et al. (2019). 'Age and gender recognition from speech signals'. In: *Journal of Physics: Conference Series* Vol. 1410, nr. 1, p. 012073). IOP Publishing.

Alzantot, M., B. Balaji & M. Srivastava (2017). 'Did you hear that? Adversarial Examples Against Automatic Speech Recognition'. In: *31st Conference on Neural Information Processing Systems*.

Amazon (2020). 'Alexa for Hospitality'. <https://www.amazon.com/alexahospitality>.

Amazon (2020). 'Alexa Priza. Amazon'. <https://developer.amazon.com/en-US/alexa/alexa-startups/alexa-fund>

Amnesty International. (2019). *Surveillance Giants: How The Business Model Of Google And Facebook Threatens Human Rights*. Amnesty International.
https://www.amnesty.nl/content/uploads/2019/11/20191119_FINAL_Surveillance-giants-online-FINAL.pdf?x68150

Anand, P. 'The Reality Behind Voice Shopping Hype'. In: *The Information*, 6 Augustus 2018. <https://www.theinformation.com/articles/the-reality-behind-voice-shopping-hype>

Arcon, N., P. Klein & J. Dombroski (2017). 'Effects of Dictation, Speech to Text, and Handwriting on the Written Composition of Elementary School English Language Learners'. In: *Reading & Writing Quarterly*, 33, nr. 6, pp. 533-548.

Arik, S., et al. (2018). 'Neural voice cloning with a few samples'. In: *Advances in Neural Information Processing Systems*. pp. 10019-10029.
<https://arxiv.org/pdf/1802.06006.pdf>

Asimov, I. (2008). *I Robot*. New York City: Penguin Random House.

Baig, E. 'Say thank you and please: Should you be polite with Alexa and the Google Assistant?'. In: *USA Today*, 10 oktober 2019.
<https://eu.usatoday.com/story/tech/2019/10/10/do-ai-driven-voice-assistants-we-increasingly-rely-weather-news-homework-help-otherwise-keep-us-info/3928733002/>

Baraniuk, C. 'How talking machines are taking call centre jobs'. In: *BBC News*, 24 augustus 2018. <https://www.bbc.com/news/business-45272835>

Barnes, E, N. Fandos & D. Hakim. 'White House Ukraine Expert Sought to Correct Transcript of Trump Call'. In: *The New York Times*, 19 november 2019.
<https://www.nytimes.com/2019/10/29/us/politics/alexander-vindman-trump-ukraine.html>

- Barocas, S. & Selbst, A. D. (2016). 'Big data's disparate impact'. In: *104 California Law Review* 671.
- Bharadwaj, R. 'Voice and Speech Recognition in Banking – What's Possible Today'. In: *Emerj*, 11 juli 2019. <https://emerj.com/ai-sector-overviews/voice-speech-recognition-banking/>
- Bremmer, D. 'Slimme spraakassistent houdt ouderen langer zelfstandig'. In: *Algemeen Dagblad*, 3 april 2019. <https://www.ad.nl/tech/slimme-spraakassistent-houdt-ouderen-langer-zelfstandig~af0a3bea/?referrer=https://www.google.com/>
- Carmen, A. 'Volkswagen now lets Apple users unlock their cars with Siri'. In: *The Verge*, 12 November 2018. <https://www.theverge.com/2018/11/12/18087416/volkswagen-vw-car-net-app-siri-shortcuts>
- CBinsights (2019). 'How Big Tech Is Battling To Own The \$49B Voice Market'. In: *CBinsights Research*, 13 februari 2019. <https://www.cbinsights.com/research/facebook-amazon-microsoft-google-apple-voice/>
- Chaudhari, S. & R. Kagalkar. (2014) 'A review of automatic speaker age classification, recognition and identifying speaker emotion using voice signal.' In: *International Journal of Science and Research* 3 nr. 11 pp. 1307-1311.
- Chen, A. 'Why companies want to mine the secrets in your voice'. In: *The Verge*, 14 maart 2019. <https://www.theverge.com/2019/3/14/18264458/voice-technology-speech-analysis-mental-health-risk-privacy>
- Chiu, C. et al. (2018). 'State-of-the-art speech recognition with sequence-to-sequence models'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4774-4778). IEEE.
- Chowdhury, G. (2003). 'Natural language processing'. In: *Annual review of information science and technology* 37 nr. 1, pp. 51-89.
- Conner, K. '9 best Alexa games to play on your Amazon Echo'. In: *CNET*, 5 maart 2020. <https://www.cnet.com/how-to/9-best-alexa-games-to-play-on-your-amazon-echo/>
- Cox, T. (2019). *Now You're Talking. The story of human conversation from the Neanderthals to Artificial Intelligence*. Penguin.
- Cox, T. (2020). Siri and Alexa Fails: Frustrations With Voice Search. The Manifest. <https://themanifest.com/digital-marketing/resources/siri-alexa-fails-frustrations-with-voice-search>
- Danaher, J. (2018). 'Toward an Ethics of AI Assistants: an Initial Framework'. In: *Philosophy & Technology*, 31, pp. 629–653.
- Day, M., G. Turner, & N. Drozdiak. 'Amazon Workers Are Listening to What You Tell Alexa'. In: *Bloomberg*, 11 april 2019. <https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio>

Dawar, N. 'Marketing in the Age of Alexa'. In: *Harvard Business Publishing*, mei 2018. <https://hbr.org/2018/05/marketing-in-the-age-of-alexa>

Dujardin, A. 'Aan het eind van deze eeuw is het aantal talen op de wereld gehalveerd'. In: *Trouw*, 30 april 2017.

Dzieza, J. How Hard Will The Robots Make Us Work? In: *The Verge*, 27 februari, 2020. <https://www.theverge.com/2020/2/27/21155254/automation-robots-unemployment-jobs-vs-human-google-amazon>

Europese Commissie (EC) (2020). *Witboek. over kunstmatige intelligentie - een Europese benadering op basis van excellentie en vertrouwen*. Brussel. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_nl.pdf

Faulkner, C. 'Issa Rae is the next celebrity voice coming to Google Assistant'. In: *The Verge*, 10 oktober 2019. <https://www.theverge.com/2019/10/10/20906823/issa-rae-google-assistant-celebrity-voice-cameo-android-ios-nest-hub-home-wavenet>.

French, K. 'Something in Your Voice'. In: *Proto*, 7 oktober 2019. <http://protomag.com/articles/something-your-voice>

Fussel, S. 'Alexa Wants to Know How You're Feeling Today'. In: *The Atlantic*, 12 oktober 2018. <https://www.theatlantic.com/technology/archive/2018/10/alexa-emotion-detection-ai-surveillance/572884/>

Gershgorn, D. 'Here's How Amazon Alexa Will Recognize When You're Frustrated'. In: *Medium*, 27 september 2019.

Glynn, F. 'What is voice picking?' In: 6River systems 18 januari 2020. <https://6river.com/what-is-voice-picking/>

Gupta, M., S. Bharti & Agarwal, S. (2019). 'Gender-based speaker recognition from speech signals using GMM model'. In: *Modern Physics Letters B*, 33 nr. 35, 1950438.

Habermas, J. (1997). *Between Facts and Norms*. Polity Press.

Hager, R. 'Google confirms new voice-confirmation feature for purchases in Assistant'. In: *Android Police*, 25 mei 2020. <https://www.androidpolice.com/2020/05/25/google-assistant-gets-new-confirm-with-voice-match-setting-for-payments/>

Hamer, J. et al. (2019). *Cyberspace zonder conflict. Op zoek naar de-escalatie van het internationale informatieconflict*. Den Haag: Rathenau Instituut.

Harris, D. 'Baidu explains how it's mastering Mandarin with deep learning'. In: *Medium*, 11 augustus 2015. <https://medium.com/s-c-a-l-e/how-baidu-mastered-mandarin-with-deep-learning-and-lots-of-data-1d94032564a5>

Harris, J. 'Voice chat with Replika, a socialbot'. In: *Medium*, 7 januari 2019. <https://medium.com/speaking-naturally/voice-chat-with-replika-a-socialbot-a9d4bcaacea8>

Henzi, R. & O. Wright. '3 ways voice technology will change your life'. In: *World Economic Forum*, 11 Juni 2019. <https://www.weforum.org/agenda/2019/06/how-voice-technology-will-change-your-life/>

Hern, A. 'Apple contractors 'regularly hear confidential details' on Siri recordings'. In: *The Guardian*, 26 juli 2019. <https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings>

Hirschberg, J., & C. D. Manning. (2015). 'Advances in natural language processing'. In: *Science* 349 nr. 6245, pp.261-266.

Hoy, M. B. (2018). 'Alexa, siri, cortana, and more: An introduction to voice assistants'. In: *Medical reference services quarterly*, 37 nr. 1, pp. 81-88

Howard, A & J. Borenstein (2017). 'The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity'. In: *Science and Engineering Ethics*, 24, pp. 1521–1536.

Huang, X., J. Baker & R. Reddy (2014). 'A historical perspective of speech recognition'. In: *Communications of the ACM*, 57 nr. 1, pp. 94-103.

iTalk2Learn (2015). Website: <https://www.italk2learn.com/>.

Jacobs, A. 'Siri in het EPD: heeft spraakbesturing ook toekomst in de zorg?' In: *Smarthealth*, 23 augustus 2018. <https://www.smarthealth.nl/2018/08/23/siri-in-het-epd-heeft-spraakbesturing-ook-toekomst-van-de-zorg/>

Jawad, U. 'IBM beats Microsoft's word error rate in speech recognition, achieves 5.5%'. In: *Neowin*. 11 maart 2017. <https://www.neowin.net/news/ibm-beats-microsofts-word-error-rate-in-speech-recognition-achieves-55>

Jothilakshmi, S. & V. N. Gudivada. (2016). 'Large Scale Data Enabled Evolution of Spoken Language Research and Applications'. In: *Handbook of Statistics* 35, pp. 301-340

Jyothi, P. *Automatic Speech Recognition - An Overview* [video]. Microsoft Research, 11 September 2017. <https://www.youtube.com/watch?v=q67z7PTGRi8>

Jyothi, P. *State-of-the-Art in Speech Technologies* [video]. Academic Research Summit 2018, 24 januari, 2018. <https://www.youtube.com/watch?v=AHk51EsRlqg>

Kamp, R. 'Smartphone gebruiken in de auto'. In: Consumentenbond: 28 Maart 2019. <https://www.consumentenbond.nl/smartphone/smartphone-in-de-auto>

Kantar. 'Gebruik smart speakers groeit explosief'. In: TNS-NIPO, 14 maart 2019. <http://www.tns-nipo.com/nieuws/persberichten/gebruik-smart-speakers-groeit-explosief?lang=nl-NL>

Kartalidis, N. (2018). *Speech recognition in construction equipment: Creating a voice assistant for an autonomous wheel loader*. Uppsala Universitet. <https://pdfs.semanticscholar.org/c222/ba625f848912abf524c042e88ddc499a659e.pdf>

- Kepuska, V. G. & Bohouta (2018). 'Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* pp. 99-103
- Kimmich, M. (2019). *Smart Homes: Strategic Opportunities, Business Models & Competitive Landscape 2019-2024*. Hampshire: Juniper Research.
- Kinsella, B. 'Amazon Files for Patent to Detect User Illness and Emotional State by Analyzing Voice Data'. In: Voicebot.ai, 10 oktober 2018.
- Kinsella, B. & A. Mutchler (2018). *Smart speaker consumer adoption report march 2018*. U.S: Voicebot. <https://voicebot.ai/wp-content/uploads/2018/10/voicebot-smart-speaker-consumer-adoption-report.pdf>
- Kinsella, B. (2019). *U.S. Smartspeaker consumer adoption report 2019*. U.S: Voicebot. <https://voicebot.ai/smart-speaker-consumer-adoption-report-2019/>
- Kinsella, B. 'Why Tech Giants Are So Desperate to Provide Your Voice Assistant'. In: *Harvard Business Review*, 7 mei 2019. <https://hbr.org/2019/05/why-tech-giants-are-so-desperate-to-provide-your-voice-assistant>
- Kishore, S. P., & A. W. Black (2003). 'Unit size in unit selection speech synthesis'. In: *Eighth European Conference on Speech Communication and Technology*. pp. 1317-1320.
- Kodish-Wachs, J. et al. (2018). 'A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech'. In: *AMIA Symposium*, pp. 683–689.
- Kofman, A. Interpol rolls out international voice identification database using samples from 192 law enforcement agencies. In: *The Intercept*, 25 juni 2018. <https://theintercept.com/2018/06/25/interpol-voice-identification-database/>
- Kohn, M. 'Is the era of artificial speech translation upon us?' In: *The Guardian*. 17 februari 2019. <https://www.theguardian.com/technology/2019/feb/17/is-the-era-of-artificial-speech-translation-upon-us>
- Koksal, I. (2018). 'Voice-First Devices Are The Next Big Thing -- Here's Why'. In: *Forbes* 1 februari 2018. <https://www.forbes.com/sites/ilkerkoksal/2018/02/01/voice-first-devices-are-the-next-big-thing-heres-why/#361e22f36873>
- Kool, L., et al. (2017). *Opwaarderen. Borgen van publieke waarden in de digitale samenleving*. Den Haag: Rathenau Instituut.
- Kool, L., J. Timmer. & R. van Est., (2014). *Eerlijk Advies. De opkomst van de e-coach*. Den Haag: Rathenau Instituut.
- Kool, L. & R. van Est. (2015). *Werken aan de robotsamenleving: Visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid*. Den Haag: Rathenau Instituut.
- Koops, B. J. (2019). *Privacyconcepten voor de 21e eeuw*. Nijmegen: Ars Aequi, 68 nr. 7/8, pp. 532-544.

Kretzschmar, K et al. (2019). 'Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support'. In: *Biomedical Informatics Insights*, 11, pp. 1–9.

Kritzler, M., et al. (2019). 'Digital Companion for Industry'. In: *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 663-667

Kuligowska, K., P. Kisielewicz & A. Włodarz. (2018). 'Speech synthesis systems: disadvantages and limitations'. In: *International Journal of Engineering & Technology* 7, pp. 234-239.

Langley, H. & J. P. Tuohy. 'Smart home privacy: What Amazon, Google and Apple do with your data'. In: *The Ambient*, 8 november 2019. <https://www.the-ambient.com/features/how-amazon-google-apple-use-smart-speaker-data-338>

Leswing, K. 'People found a really easy way to make Siri curse'. In: *Business Insider*, 30 april 2018. <https://www.businessinsider.nl/apple-siri-swears-when-asked-for-second-definition-of-mother-2018-4/>

Leventon, W. 'Alexa and Siri, meet Athena for machining'. In: *Cutting Tool Engineering*, 8 januari 2019.

Lens-FitzGerald, e.a. (2020). *Voice Verbindt: Slimme spraaktechnologie verbetert leefkwaliteit van ouderen*. Projectzilver.

<https://projectzilver.com/onderzoeksrapport/> Li, B., et al. (2020). 'Towards fast and accurate streaming end-to-end ASR'. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6069-6073

Lieberman, H. et al. (2005). 'How to wreck a nice beach you sing calm incense'. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 278-280.

Liu, K. et al. (2019). *A review of the literature on computerized speech-to-text accommodations (NCEO Report 414)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

MacCartney, B. 'Computational Linguistics (aka Natural Language Processing)' [Presentation]. Stanford University, 26 mei 2011. <https://nlp.stanford.edu/~wcmac/papers/20110526-symsys-100-nlp.pdf>

Manning, C. 'Lecture 1 | Natural Language Processing with Deep Learning' [video]. Stanford University School of Engineering, 3 april 2017. https://www.youtube.com/watch?v=OQQ-W_63UgQ

Mattin, D. 'Voice technology could soon be your new best friend. Here's why'. In: *World Economic Forum*, 10 Juli 2019. <https://www.weforum.org/agenda/2019/07/voice-technology-personalization/>

Marmar, C. R., et al. (2019). 'Speech-based markers for posttraumatic stress disorder in US veterans'. In: *Depression and anxiety*, 36 nr. 7, pp. 607-616.

Meeker, M. (2018). *INTERNET TRENDS 2018*. California: Kleiner Perkins. <https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/>

Meiling, C. 'Demand for AI virtual anchors set to increase'. In: *China Daily*, 14 juni 2019.

<http://www.chinadaily.com.cn/a/201906/14/WS5d03013aa3103dbf14328348.html>

Meridith, S. 'A new era in commerce': Amazon Pay VP says voice payments potential is 'phenomenal'. In: *CNBC*, 4 juni 2019.

Meyer, D. 'Introducing Alexa Conversations (Preview), a New AI-Driven Approach to Natural Dialogs through the Alexa Skills Kit'. In: Amazon Alexa, 5 juni 2019.

<https://developer.amazon.com/blogs/alexa/post/44499221-01ff-460a-a9ee-d4e9198ef98d/introducing-alexa-conversations-preview>

Mittal, U. (2019). 'Design of data virtual assistant using natural language processing'. In: *Journal of the Gujarat Research Society*, 21 nr. 8, pp. 786-791.

Muller, J. Improved voice controls could increase car safety. In: *Axios*, 29 mei 2019.

<https://www.axios.com/autonomous-vehicles-safety-voice-controls-ec6845d5-3480-4ed4-a15b-cd5641130982.html>

Müller, C. (2006). 'Automatic recognition of speakers' age and gender on the basis of empirical studies'. In: *Ninth International Conference on Spoken Language Processing*.

Multiscope. 'Smart home markt groeit tot 2,5 miljard euro'. In: *Multiscope*, 12 maart 2020 <http://www.multiscope.nl/persberichten/smart-home-markt-groeit.html>

Mwiti, D. A. 'Guide for Automatic Speech Recognition'. In: *Medium / Heartbeat*, 4 September 2019. <https://heartbeat.fritz.ai/a-2019-guide-for-automatic-speech-recognition-f1e1129a141c>

Nass, C. I., & S. Brave. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT press.

Nederlandse Publieke Omroep (NPO) (2020). *De 10 lessen uit 15 smart speaker projecten*. <https://innovatie.npo.nl/projecten/de-10-lessen-uit-15-smart-speaker-projecten>

Newman, J. 'Can Mycroft's Privacy-Centric Voice Assistant Take On Alexa and Google?' In: *Fastcompany*. 29 januari 2018.

<https://www.fastcompany.com/40522226/can-mycrofts-privacy-centric-voice-assistant-take-on-alexa-and-google>

Olson C. & K. Kemery. (2019). *Voice Report*. Microsoft.

<https://about.ads.microsoft.com/en-us/insights/2019-voice-report>

Oord, A. V. D. et al. (2016). 'Wavenet: A generative model for raw audio'. London:

Google DeepMind. <https://arxiv.org/pdf/1609.03499.pdf>

O'Shaughnessy, D. (2013). 'Acoustic analysis for automatic speech recognition'. In: *Proceedings of the IEEE 101*, nr. 5, pp. 1038-1053.

Pardes, A. 'Hey Alexa, Why Is Voice Shopping So Lousy?' In: *WIRED*, 17 Juni 2019. <https://www.wired.com/story/why-is-voice-shopping-bad/>

Pichai, S. (2018) Keynote (Google I/O '18). Google Developers.
<https://www.youtube.com/watch?v=ogfYd705cRs&t=1014s>

Pichai, S. (2019) Google Keynote (Google I/O '19). Google Developers.
<https://www.youtube.com/watch?v=lyRPyRKHO8M>

Pisanski, K., et al. (2014). 'Vocal indicators of body size in men and women: a meta-analysis'. In: *Animal Behaviour* 95, pp. 89-99.

Place, S., Blanch-Hartigan, D., Rubin, C., Gorrostieta, C., Mead, C., Kane, J., ... & Azarbajani, A. (2017). Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of medical Internet research*, 19(3), e75. Protalinski, E. (2019). ProBeat: Has Google's word error rate progress stalled?. In: Venturebeat.
<https://venturebeat.com/2019/05/10/probeat-has-googles-word-error-rate-progress-stalled/>

Pundak, G., et al. (2018). 'Deep context: end-to-end contextual speech recognition'. In: *2018 IEEE Spoken Language Technology Workshop (SLT)* pp. 418-425.

Rathenau Instituut (2020). Reactie op verzamelwet gegevensbescherming. Brief aan Minister Dekker, 14 juli 2020

Reilly, M.. 'Should an Amazon Echo Help Solve a Murder?'. In: *MIT Technology Review*, 27 december 2016.
<https://www.technologyreview.com/2016/12/27/154864/should-an-amazon-echo-help-solve-a-murder/>

Reuters. 'Germany planning to access voice assistant data to tackle crime'. In: *Reuters*, 5 juni 2019. <https://www.reuters.com/article/uk-germany-crime-alexa/germany-planning-to-access-voice-assistant-data-to-tackle-crime-idUKKCN1T61IE>

RTL Nieuws. 'Grote zorgen oorartsen: jongeren vaker met gehoorproblemen door koptelefoons en oortjes'. In: *RTL Nieuws*, 23 augustus 2019.
<https://www.rtlnieuws.nl/nieuws/nederland/artikel/4822536/gehoorschade-oren-knoarts-dokter-tinnitus-piep-oorsuizen-muziek>

Ruane, E., A. Birhane & A. Ventresque (2019). 'Conversational AI: Social and Ethical Considerations'. In: *AICS - 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*.

Sankaranarayanan, P. (2017). *New Zealand Government Case Study*. Auraya Systems. <https://aurayasystems.com/2017/04/28/new-zealand-government-case-study/>

Sawers, P. 'How Duolingo is using AI to humanize virtual language lessons'. In: *Venture Beat*, 5 juli 2019. <https://venturebeat.com/2019/07/05/how-duolingo-is-using-ai-to-humanize-virtual-language-lessons/>

Schwartz, E. 'New Resemble AI Software Turns 3-Minute Records into Synthetic Speech Profiles'. In: *Voicebot.ai*, 30 december 2019.

<https://voicebot.ai/2019/12/30/new-resemble-ai-software-turns-3-minute-records-into-synthetic-speech-profiles/>

Sedaaghi, M. (2009). 'A Comparative Study of Gender and Age Classification in Speech Signals'. In: *Iranian Journal of Electrical & Electronic Engineering* 5.

Senseable Intelligence Group. (2019). Senseable Intelligence Group: Research. Cambridge: McGovern Institute for Brain Research.
<https://satra.cogitatum.org/group/research/>

Shank, D., C. et al. (2019). 'Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence'. In: *Computers in Human Behavior* 98, pp. 256-266.

Shulevitz, J. 'Alexa, Should We Trust You?' In: *The Atlantic*. November 2018.
<https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/>

Simon, M. "The Genderless Digital Voice the World Needs Right Now". In: *Wired*. 3 november 2019.

Simonite, T. 'This Call May Be Monitored for Tone and Emotion'. In: *WIRED*. 19 maart 2018. <https://www.wired.com/story/this-call-may-be-monitored-for-tone-and-emotion/>

Smith, C. 'Alexa and Siri Can Hear This Hidden Command. You Can't'. In: *The New York Times*, 10 mei 2018.

Softengi (2020). 'Voice Recognition Technology Is a Must in Any Risky Industry'. Website: Softengi. <https://softengi.com/blog/voice-recognition-technology-is-a-must-in-any-risky-industry/>

Sokol, N., W. Chen & B. Donmez. (2017). 'Voice-Controlled In-Vehicle Systems: Effects of Voice-Recognition Accuracy in the Presence of Background Noise'. In: *Conference: Driving Assessment Conference*. pp. 158-164

Sondhi, S., et al. (2015). 'Vocal indicators of emotional stress'. In: *International Journal of Computer Applications*, 122, nr. 15. pp. 38-43

Staff, I. 'By voice or location, Israeli apps can determine your risk of coronavirus'. In: *The Times of Israel*, 31 maart 2020. <https://www.timesofisrael.com/by-voice-or-location-israeli-apps-can-determine-your-risk-of-coronavirus/>

Staff, I. 'Israeli startup aims to identify coronavirus carriers using their voice'. In: *The Times of Israel*, 25 maart 2020. <https://www.timesofisrael.com/israeli-startup-aims-to-identify-coronavirus-carriers-using-their-voice/>

Statt, N. "Amazon sent 1,700 Alexa voice recordings to the wrong user following data request". In: *The Verge*, 20 december 2018.
<https://www.theverge.com/2018/12/20/18150531/amazon-alexa-voice-recordings-wrong-user-gdpr-privacy-ai>

Strayer, D & J. Cooper. (2015). *Up to 27 seconds of inattention after talking to your car or smartphone*. University of Utah. <https://unews.utah.edu/up-to-27-seconds-of->

[inattention-after-talking-to-your-car-or-smart-phone/?doing_wp_cron=1586506141.6724519729614257812500](#)

Stucke, M. & A. Ezrachi. 'The Subtle Ways Your Digital Assistant Might Manipulate You'. In: *Wired*, 29 november 2016.

Stinson, L. 'The Surprising Repercussions of Making AI Assistants Sound Human'. In: *WIRED*, 12 mei 2017. <https://www.wired.com/2017/05/surprising-repercussions-making-ai-assistants-sound-human/>

Streijl, R.C., S. Winkler, & D. Hands. (2014) 'Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives'. In: *Multimedia Systems*, pp.213-227

Sundermeyer, M., R. Schlüter, & H. Ney (2012). 'LSTM neural networks for language modeling'. In: *INTERSPEECH-2012*, pp. 194-197

Synnaeve, G. et al., (2020). 'End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures'. In: *The 37th International Conference on Machine Learning*.

Szymkowski, S. 'Apple's self-driving car system could use voice, gesture guidance'. In: *Roadshow*, 27 januari 2020. <https://www.cnet.com/roadshow/news/apple-self-driving-car-system-voice-gesture-guidance/>

Tatman, R. (2017). 'Gender and Dialect Bias in YouTube's Automatic Captions'. In: *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pp. 53–59.

Tokuday, K., & H. Zen. (2015). 'Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis'. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 4215-4219

Turkle, S. (2017). *Alone Together: Why We Expect More from Technology and Less from Each Other (Third Edition)*. U.S: Ingram Publisher Services

Turner, M. 'Smart Home 101: How to Develop for the Connected Home (Google I/O'19)' [video]. *Google Developers*. 8 mei 2019. <https://www.youtube.com/watch?v=SJ2KYVKfURA>

UNESCO (2019). *I'd blush if I could*. Equals skill coalition & Unesco. <https://en.unesco.org/ld-blush-if-i-could>

Van Bergen, W & Van Gelder, H. 'Shoppen via de slimme speaker'. In: *De Telegraaf*: 10 oktober 2018. <https://www.telegraaf.nl/financieel/2659211/shoppen-via-slimme-speaker>

Van Camp, J. 'My Jibo Is Dying and It's Breaking My Heart'. In: *WIRED*, 3 augustus 2019. <https://www.wired.com/story/jibo-is-dying-eulogy/>

Van Dijck, J., T. Poell & M. De Waal (2016). *De Platformsamenleving. Strijd om publieke waarden in een online wereld*. Amsterdam: Amsterdam University Press.

Van Est, R., J. Gerritsen, & L. Kool (2017). *Human rights in the robot age. Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality*. Den Haag: Rathenau Instituut.

Van Keulen, I. et al. (2018). *Digitalisering van het nieuws. Online nieuwsgedrag, desinformatie en personalisatie in Nederland*. Den Haag: Rathenau Instituut.

Vernuccio, M., M. Patrizi & A. Pastore (2020). 'Developing voice-based branding: insights from the Mercedes case'. In: *Journal of Product & Brand Management*.

Vermeend, W. & J. W. Timmer (2016). *Internet of things. De nieuwste internet megatrend met een enorme impact op economie werken ondernemen*. Den Haag: Einstein Books en Ebooks

Vincent, J. 'Lyrebird claims it can recreate any voice using just one minute of sample audio.' In: *The Verge*, 24 april 2017.
<https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird>

Vlahos, J. 'Amazon Alexa and the Search for the One Perfect Answer'. In: *WIRED*, 18 februari, 2019. <https://www.wired.com/story/amazon-alexa-search-for-the-one-perfect-answer/>

Vlahos, J. 'Barbie wants to get to know your child'. In: *The New York Times*, 16 september 2015. <https://www.nytimes.com/2015/09/20/magazine/barbie-wants-to-get-to-know-your-child.html>

Waarlo, N. "Het eerste woordje van de baby: 'mama' of 'papa'? Straks blijkt het 'Siri' of 'Alexa'". In: *De Volkskrant*, 9 augustus 2019.
<https://www.volkskrant.nl/wetenschap/het-eerste-woordje-van-de-baby-mama-of-papa-straks-blijkt-het-siri-of-alexa>

Warren, T. 'Microsoft no longer sees Cortana as an Alexa or Google Assistant competitor'. In: *The Verge*, 18 januari, 2018.
<https://www.theverge.com/2019/1/18/18187992/microsoft-cortana-satya-nadella-alexa-google-assistant-competitor>

Weizenbaum, J. (1966). 'ELIZA---a computer program for the study of natural language communication between man and machine'. In: *Communications of the ACM*, 9 nr. 1, pp. 36-45.

Welch, C. 'Waze now lets you record your own turn-by-turn audio'. In: *The Verge*. 8 mei 2017. <https://www.theverge.com/2017/5/8/15585638/waze-record-custom-navigation-audio>

Wenderow, T. 'Vocalis CEO Tal Wenderow Discusses Vocal Biomarkers, Healthcare, and Coronavirus – Voicebot Podcast Ep 145' [audio]. *Voicebot.ai*, 12 april 2020. <https://voicebot.ai/2020/04/12/vocalis-ceo-tal-wenderow-discusses-vocal-biomarkers-healthcare-and-coronavirus-voicebot-podcast-ep-145/>

Williams, R. 'Amazon expands Alexa with voice-powered wearables'. In: *Mobile Marketer*, 26 september 2019. <https://www.mobilemarketer.com/news/amazon-expands-alexa-with-voice-powered-wearables/563740/>

Wolters Kluwer. 'Wolters Kluwer provides voice-enabled search powered by Nuance Dragon Medical One to millions of clinicians using leading clinical decision support'. In: *Wolters Kluwer*, 17 juni 2020. <https://www.wolterskluwer.com/en/news/dragon-medical-voice-search-uptodate>

Wu, H., et al. (2020). 'Defense against adversarial attacks on spoofing countermeasures of ASV'. In: *Electrical Engineering and Systems Science*, arXiv:2003.03065.

Yoffie, D. et al. (2018). 'Voice War: Hey Google vs. Alexa vs. Siri.' In: *Harvard Business School Case Collection*, pp. 718-519

Young, T., et al. (2018). 'Recent trends in deep learning based natural language processing'. In: *IEEE Computational intelligence magazine*, 13 nr. 3, pp. 55-75.

Zen, H. Generative Model-Based Text-to-Speech Synthesis [video]. *Centre for Brain, Minds and Machines (CBMM)*, 3 februari 2018. <https://www.youtube.com/watch?v=nsrSrYtKkT8>

Zhu, Z. et al. (2015). 'Application of Speech Recognition Technology to Virtual Reality System.' In: *2015 International Industrial Informatics and Computer Engineering Conference*. Atlantis Press. <https://www.atlantispress.com/proceedings/iicec-15/16911>

© Rathenau Instituut 2020

Verveelvoudigen en/of openbaarmaking van (delen van) dit werk voor creatieve, persoonlijke of educatieve doeleinden is toegestaan, mits kopieën niet gemaakt of gebruikt worden voor commerciële doeleinden en onder voorwaarde dat de kopieën de volledige bovenstaande referentie bevatten. In alle andere gevallen mag niets uit deze uitgave worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotokopie of op welke wijze dan ook, zonder voorafgaande schriftelijke toestemming.

Open Access

Het Rathenau Instituut heeft een Open Access beleid. Rapporten, achtergrondstudies, wetenschappelijke artikelen, software worden vrij beschikbaar gepubliceerd. Onderzoeksgegevens komen beschikbaar met inachtneming van wettelijke bepalingen en ethische normen voor onderzoek over rechten van derden, privacy, en auteursrecht.

Contactgegevens

Anna van Saksenlaan 51
Postbus 95366
2509 CJ Den Haag
070-342 15 42
info@rathenau.nl
www.rathenau.nl

Bestuur van het Rathenau Instituut

Mw. Gerdi Verbeet
Prof. dr. Noelle Aarts
Drs. Felix Cohen
Dr. Hans Dröge
Dr. Laurence Guérin
Dr. Janneke Hoekstra, MSc
Prof. mr. dr. Erwin Muller
Drs. Rajash Rawal
Prof. dr. ir. Peter-Paul Verbeek
Dr. Ir. Melanie Peters – secretaris

Het Rathenau Instituut stimuleert de publieke en politieke meningsvorming over de maatschappelijke aspecten van wetenschap en technologie. We doen onderzoek en organiseren het debat over wetenschap, innovatie en nieuwe technologieën.

Rathenau Instituut