1.2 Mens-machine interactie

Dr. O.E. (Odette) Scharenborg
Computer Science
Technische Universiteit Delft

http://homepage.tudelft.nl/f7h35

- Lezing gehouden voor de Koninklijke Maatschappij voor Natuurkunde 'Diligentia' te 's-Gravenhage op 7 oktober 2019.
- Een video opname van de lezing is te zien op www.natuurwetenschappen-diligentia.nl.
- Deze jaarlijks terugkerende JONG DILIGENTIA LEZING is primair bedoeld voor scholieren.

Samenvatting van de lezing:
Computers zijn niet meer uit onze samenleving weg te denken. Vroeger werden ze vooral gebruikt voor ingewikkelde berekeningen of voor het uitwerken van verslagen; tegenwoordig kijken we naar filmpjes, kletsen we met vrienden, versturen we e-mail, en winkelen we via onze smartphones. Veel van onze interactie met computers gaat via tekst, maar vaak zou het handiger zijn om je stem te gebruiken om een berichtje op te stellen of een telefoonnummer op te vragen. Als je aan het fietsen bent bijvoorbeeld.
In deze lezing legde dr. Odette Scharenborg van de TU Delft prachtig uit hoe we kunnen communiceren met computers door middel van spraak. De eerste vraag was wat spraak eigenlijk is. Daarna besprak dr. Odette Scharenborg de basisprincipes waarmee computers spraak kunnen verstaan en hoe computers spraak kunnen genereren.

Introductie

Mens-machine interactie is belangrijk voor het bouwen van gebruiksvriendelijke systemen die functioneel en veilig te gebruiken zijn. Je kunt hierbij denken aan allerlei soorten machines, zoals bijvoorbeeld de machines die gebruikt worden om auto's aan een lopende band in elkaar te zetten, maar ook sociale robots. Sociale robots worden steeds meer gebruikt, bijvoorbeeld om gasten te ontvangen in een ziekenhuis of om eenzame ouderen gezelschap te houden. Er zijn diverse manieren om met een machine of robot te interacteren: bijvoorbeeld via een toetsenbord, een muis, joystick, knoppen, touchscreen of door er tegen te praten. In deze bijdrage gaat het over het interacteren met een machine of robot door er tegen te praten.

Er zijn meerdere goede redenen om spraak te gebruiken voor de interactie met een machine of robot. Ten eerste, er zijn veel mensen die niet goed kunnen typen of niet goed een muis kunnen gebruiken, die daar niet comfortabel mee zijn, of het door een fysieke beperking niet kunnen. Soms heb je je handen vol (bijvoorbeeld als je een grote doos draagt). Soms *mag* je niet typen, bijvoorbeeld tijdens het auto rijden of tijdens het fietsen. En de belangrijkste reden is dat spraak de meest natuurlijke vorm van communicatie is.

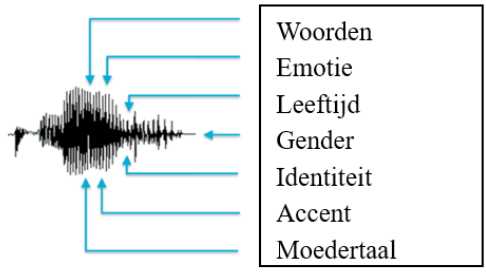
Om met een machine of robot te communiceren via spraak moeten er een aantal stappen uitgevoerd worden. Figuur 1 geeft een overzicht van deze stappen. Eerst moet het spraaksignaal omgezet worden naar tekst. Dit gebeurt in de module in figuur 1 die ASR heet, wat voor Automatic Speech Recognition (automatische spraakherkenning) staat. Deze tekst is nodig omdat de machine moet begrijpen wat de spreker bedoelt. Soms is de vraag van de spreker expliciet, bijvoorbeeld *Hoe kom ik bij zaal A53?* Soms is de vraag van de spreker impliciet: *Ik weet niet hoe ik bij zaal A53 kan komen.* De bedoeling van de spreker wordt uitgezocht door de module die NLP heet in figuur 1. NLP staat voor Natural

Language Processing (natuurlijke taalverwerking). In de volgende stap moet de benodigde informatie gezocht en gevonden worden. In de laatste stap/module wordt de gevonden informatie teruggegeven aan de gebruiker. Dit kan geschreven tekst zijn, maar ook spraak. In deze bijdrage gaan we ervan uit dat het teruggeven van de informatie ook via spraak gebeurt, en wel in de module Synt wat voor Synthese staat.

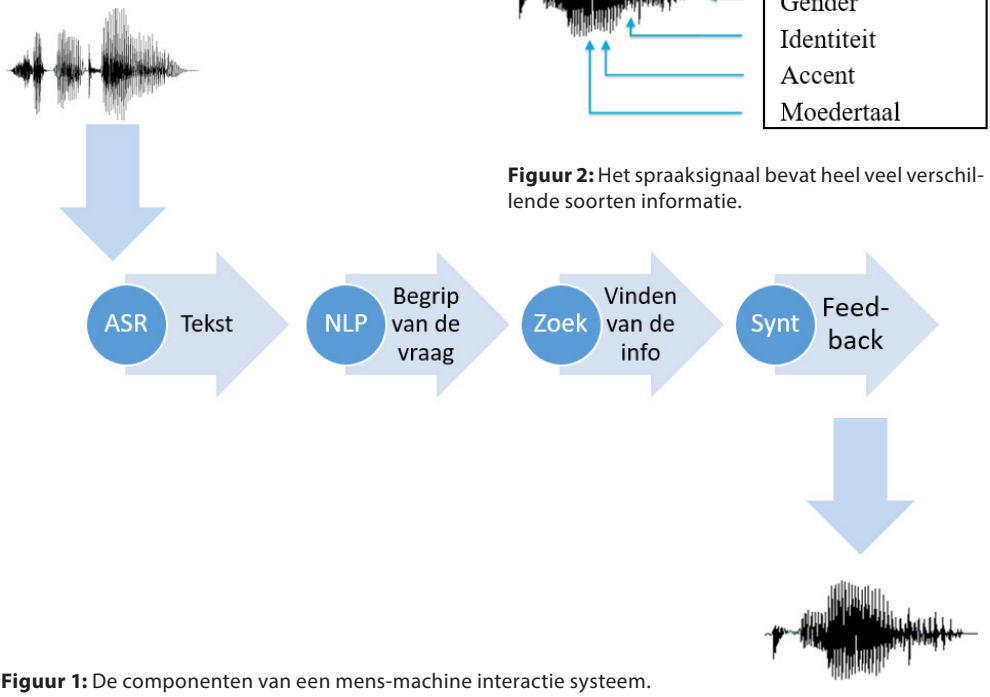
Deze bijdrage heeft als doel inzicht te verschaffen in hoe computers spraak verstaan en genereren om mens-machine/robot interactie via spraak mogelijk te maken. In de volgende vier secties leg ik eerst uit wat spraak nu eigenlijk is en hoe we het genereren. Vervolgens leg ik uit hoe automatische spraakherkenning werkt, gevolgd door een sectie over hoe goed automatische spraakherkenners nu eigenlijk zijn en hoe ze geëvalueerd worden. De laatste sectie legt uit hoe spraak geproduceerd kan worden (spraaksynthese). Is je nieuwsgierigheid geprikkeld? Lees dan meer in de twee bronnen met meer informatie aan het einde van deze bijdrage.

Wat is spraak?

Het spraaksignaal bevat heel veel informatie (zie figuur 2). Naast de woorden die de spreker heeft uitgesproken, bevat het spraaksignaal ook informatie over de emotie van de spreker, dat wil zeggen over hoe hij of zij zich voelde toen hij of zij sprak (iemand die boos is, klinkt heel anders dan iemand die verdrietig is). Ook kun je informatie uit het spraaksignaal halen over de leeftijd van de spreker, de gender van de spreker, en over *wie* er sprak, dus de identiteit van de spreker. Als iemand met een accent spreekt, kun je dit ook horen in het spraaksignaal en kun je achterhalen met welk accent de spreker spreekt. Is het een dialect of een accent omdat de spreker een taal spreekt die niet zijn of haar moedertaal is? En tot slot kun je iemands moedertaal achterhalen via het spraaksignaal, ook als de persoon niet in zijn of haar moedertaal spreekt. Om met een machine of robot te communiceren zijn we



Figuur 2: Het spraaksignaal bevat heel veel verschillende soorten informatie.



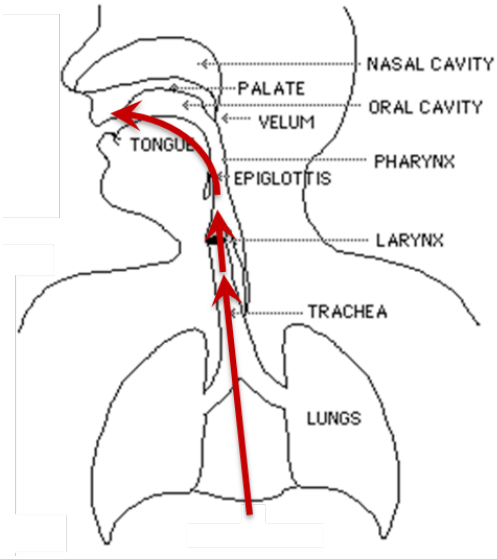
Figuur 1: De componenten van een mens-machine interactie systeem.

vooral geïnteresseerd in *wat* er gezegd is, dus in de woorden die er uitgesproken zijn.

Het maken van spraak bestaat uit drie stappen, zie figuur 3. In de eerste stap wordt er lucht uit de longen langs je stembanden (je hebt er twee) geblazen (*initiatie*). Als deze gaan trillen, heet dit *fonatie* (stap 2). Tot slot gaat de al dan niet trillende lucht door de mond- en neusholte. Hier wordt de luchtstroom vervormd, afhankelijk van de stand van de tong, de tanden, de lippen en het klepje tussen je mond- en keelholte (stap 3). Dit heet de *articulatie*-fase. Spraak bestaat dus uit klanken die gevormd worden door je mond- en keelholte van stand te veranderen. Overigens zijn geen van de lichaamsdelen die betrokken zijn bij het maken van spraak speciaal voor het spreken gemaakt. De longen zijn om adem te halen, de stembanden zijn om de luchtpijp af te kunnen sluiten als we eten of drinken, de neusholte is om de lucht te verwarmen en om te ruiken, en de tanden, tong en lippen zijn vooral bedoeld om mee te eten.

Spraakklanken

Spraakklanken worden in twee groepen verdeeld: de klinkers en de medeklinkers. Bij **klinkers** kan de luchtstroom vanuit je longen grotendeels ongehinderd door je mond naar buiten stromen. Je spraakkanaal vernauwt zich niet. Bijvoorbeeld, om een /a/[1] (als in *spr**aa**k*; vetgedrukte letters geven de klank aan) te maken, hou je je mond ver open en ligt je tong onder in je mond. De verschillende klinkers worden gemaakt doordat je lippen en tong (en onderkaak) steeds andere posities innemen.



Figuur 3: Spraak wordt geproduceerd in drie stappen: initiatie, fonatie, en articulatie.

Lippen. Voor sommige klinkers maak je een klein rondje van je lippen waarbij je je lippen ook nog iets naar buiten duwt, zoals bij de /o/ (als in *r**oo**s*). Bij de /i/ (als in *b**ie**r*) maak je je lippen juist breed en hou je ze vlak boven elkaar.

Tong. De tong kan horizontaal bewegen (van achteren naar voren). Als de tong voor in de mond een soort van verdikking maakt, wordt gezegd dat de tong voor in de mond ligt. Als de verdikking in het midden of achter in de mond ligt, wordt gezegd dat de tong in het midden dan wel achterin de mond ligt. Vergelijk maar eens de /i/ (als in *b**ie**r*; vlak achter de tanden), /ɛ/ (als in *st**e**m*; midden in de mond), /ɑ/ (als in *w**a**nd*, achter in de mond). De tong kan verder ook verticaal bewegen. Hij kan boven in de mond, onder in de mond en er tussen in liggen (ook wel de neutrale positie genoemd). Vergelijk bijvoorbeeld de ligging van de tong als je een /i/ (hoog, dicht tegen je gehemelte) of een /ɑ/ (laag, onderin je mond, als in *k**a**st*) maakt.

Medeklinkers zijn klanken waarbij je mond- of neusholte zich ergens vernauwt of helemaal afsluit. Welke klank je maakt hangt af van drie dingen: het type vernauwing/afsluiting (de 'manier van articulatie'), de plek van de vernauwing/afsluiting (de 'plaats van articulatie') en of je stembanden trillen ('stemhebbendheid').

1) Het gebruik van // om een letter of letterreeks geeft aan dat het hier niet om de letter gaat maar om de klank(reeks) die gesymboliseerd wordt met deze letter. Dit heet fonetisch schrift. Fonetisch schrift gebruikt alfabetische letters om klanken weer te geven. Dit artikel volgt het standaard fonetisch schrift van de International Phonetic Association (IPA).

Manier van articulatie. Vergelijk bijvoorbeeld de /k/ (als in ***k**ast*) met de /f/ (als in ***f**eest*). In het eerste geval sluit je je spraakkanaal heel even helemaal af door je tong achterin je mond tegen je gehemelte te drukken. Hierdoor ontstaat er druk achter je tong. Als je dan je tong weghaalt van je gehemelte hoor je een klein plofje. Dit type klank wordt daarom een 'plosief' of 'plofklank' genoemd. Bij de /f/ komen je tanden tegen je onderlip aan, maar is er geen volledige afsluiting. De lucht kan langs/door je tanden en lippen gewoon naar buiten. Je hoort hierbij een ruisje. Dit type klank wordt 'ruisklank' of 'fricatief' genoemd. Naast deze twee typen medeklinkers zijn er nog drie typen medeklinkers in het Nederlands: de nasalen (zoals de /m/ en /n/ in ***m**aan*) 'vloeiklanken' en de 'halfklinkers' (samen ook wel 'approximanten' genoemd). Bij vloeiklanken (dit zijn de /r/ zoals in ***r**echts* en de /l/ zoals in ***l**inks*) en halfklinkers (dit zijn de /j/ zoals in ***j**arig* en de /w/ zoals in ***w**ater*) is er weinig vernauwing en kan de lucht redelijk ongehinderd naar buiten stromen.

Plaats van articulatie. Je spraakkanaal kan zich op meerdere plekken vernauwen of afsluiten:

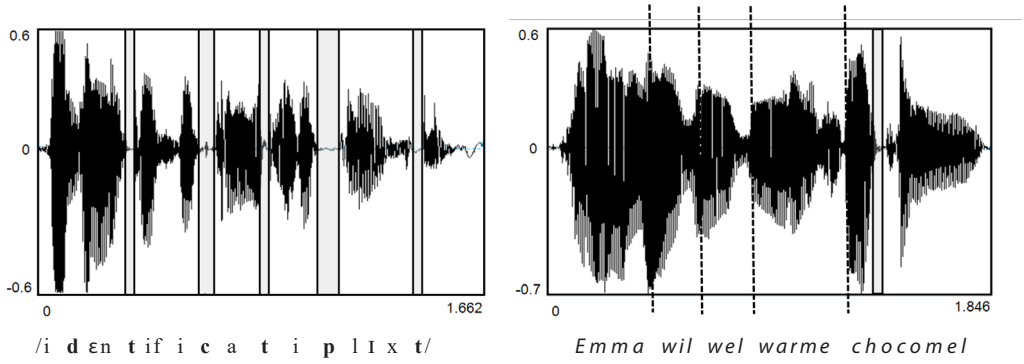
- met je lippen: bijvoorbeeld de /b/ als in ***b**ier*. Dit heet een 'bilabiale' klank.
- met je tanden op je onderlip: bijvoorbeeld de eerdergenoemde /f/. Dit heet een 'labiodentale' klank.
- met je tong net achter je bovenste tanden op de tandkas: bijvoorbeeld de /d/ als in ***d**ier*. Dit heet een 'alveolaire' klank.
- met je tong tegen verschillende plekken van je gehemelte: bijvoorbeeld de /k/ als ***k**ier* (dit heet een 'palatale' klank), en nog verder naar achteren zoals de /x/ in ***g**roen* in sommige dialecten in het oosten en het noorden van Nederland; dit heet een 'velaire' klank).

Stemhebbendheid. Als je stembanden gaan trillen door de langsstromende lucht, dan noem je dat 'stemhebbende' klanken. Trillen je stembanden niet dan zijn de klanken 'stemloos'.

Drie bijzonderheden van spraak

Woorden bestaan dus uit opeenvolgende spraakklanken en opeenvolgende woorden maken samen zinnen. Er zijn drie aspecten van spraak die belangrijk zijn om te begrijpen waarom het herkennen van spraak nog niet zo simpel is als het lijkt:

Het spraaksignaal is een **continu** signaal (het ene woord gaat in het volgende over) en er zijn **geen duidelijke grenzen** tussen woorden. Als je naar spraak luistert, moet je zelf achterhalen waar de grenzen tussen de woorden zitten. Figuur 4 laat dit met twee voorbeelden zien. Ieder paneel laat een golfvorm zien die een visuele weergave is van een spraaksignaal. Op de horizontale as staat de tijd. De klank/het woord dat als eerste uitgesproken is, staat helemaal links. Wat als laatste uitgesproken is, staat helemaal rechts. De verticale as geeft



Figuur 4: Golfvormen met de plofklanken aangegeven in de grijze balken en woordgrenzen aangegeven met de stippellijnen.

de amplitude weer, oftewel hoe hard iets klinkt. Een harde spraak(klank) heeft een grotere amplitude, dus een grotere afwijking van de nul-lijn, dan zachtere spraak(klanken). De spraaksignalen in de plaatjes zijn uitgesproken door een Nederlandse jongeman met een normale spreeksnelheid.

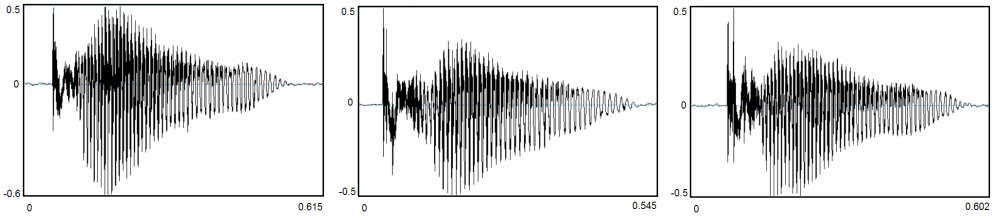
Het linkerplaatje bevat de golfvorm van één woord: *identificatieplicht*. Het rechterplaatje bevat de golfvorm van de zin: *Emma wil wel warme chocomel*. Ondanks dat het linkerplaatje maar één woord bevat, zijn er veel meer stiltes of bijna-stiltes te zien dan in het rechterplaatje dat vijf woorden bevat. Hoe kan dit? Dit komt doordat de stiltes niet veroorzaakt worden door stiltes *tussen* de woorden maar door de eerder genoemde plosieven, de klanken waarbij het spraakkanaal even helemaal dicht gaat voordat de lucht met een knalletje vrij komt. Tijdens de periode dat het spraakkanaal even helemaal dicht is, is de amplitude nagenoeg nul. Het woord in het linkerplaatje bevat veel plosieven (aangegeven met de grijze balken) terwijl het rechterplaatje maar éé plosief heeft. In het rechterplaatje is duidelijk te zien dat het éne woord naadloos overgaat in het volgende woord.

Het spraaksignaal is variabel. Iedere keer dat je een woord of zin uitspreekt klinkt het anders. Figuur 5 laat dit zien aan de hand van een voorbeeld. Alle drie de plaatjes bevatten de golfvorm van één woord. Zoals je ziet zijn er verschillen tussen de golfvormen in alle plaatjes, maar toch is het woord in het rechterplaatje *traan* en het woord in de andere twee plaatjes *trein*.

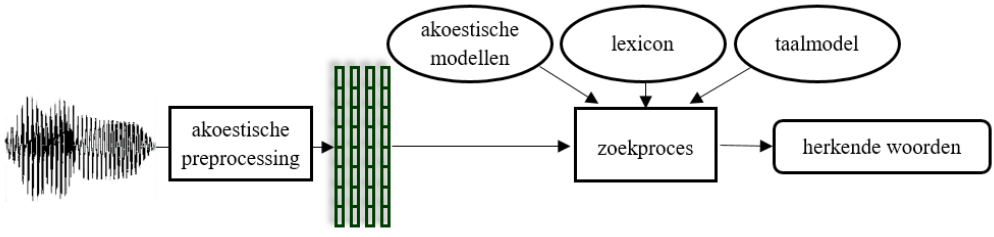
Het herkennen van spraak door computers: Automatische spraakherkenning

Een automatische spraakherkenner zet het spraaksignaal om in een reeks woorden. Het doel is om hierbij zo min mogelijk fouten te maken. Er zijn drie belangrijke informatiebronnen nodig voor het herkennen van spraak door een computer: de woordenlijst (of het lexicon), de akoestische modellen en het taalmodel (zie figuur 6).

In de eerste stap wordt het spraaksignaal omgezet naar numerieke *vectoren*. Dit gebeurt in de *akoestische preprocessing* fase. Deze vectoren bevatten alleen die informatie die nodig is voor het herkennen van de spraak.



Figuur 5: Ieder plaatje bevat de golfvorm van één uitgesproken Nederlands woord. Welke golfvorm is van een ander woord?



Figuur 6: De bouwstenen van een automatisch spraakherkenningssysteem.

De woordenlijst. Iedere automatische spraakherkenner bevat een woordenlijst. Daarin staan alle woorden die de automatische spraakherkenner kan herkennen. Dit betekent dus ook dat als een woord niet in het lexicon staat, dit woord niet herkend kan worden. Bij het maken van een toepassing voor een automatische spraakherkenner is het dus cruciaal dat alle woorden die de spraakherkenner moet kunnen herkennen ook daadwerkelijk in het lexicon staan. Van elk woord in het lexicon is er ook een omschrijving in termen van een beperkte set van segmenten (meestal zijn dit klanken) waaruit het woord opgebouwd is (zie ook 'akoestisch model' hieronder). Een belangrijk verschil tussen een mens en een computer is overigens dat een automatische spraakherkenner niet kan aangeven dat het een woord niet kent. Wij kunnen dat wel.

Akoestisch model. Een akoestisch model is een soort van gemiddelde representatie van het spraaksignaal van een segment en wordt getraind met heel veel voorbeelden van dat segment. In de meeste automatische spraakherkenners is zo'n segment een klank (zie hierboven). Om een woord te herkennen, herkent een automatische spraakherkenner reeksen van deze klanken. Een automatische spraakherkenner heeft dus in principe van alle klanken van een taal een akoestisch model. Het lexicon bevat dan een omschrijving van het woord in termen van die akoestische modellen klanken. De woorden in het lexicon bepalen dan welke volgordes van akoestische modellen geanalyseerd worden tijdens het spraakherkenningsproces. Tijdens het spraakherkenningsproces vergelijkt de spraakherkenner het spraaksignaal met zijn akoestische modellen en berekent op welke akoestische modellen het spraaksignaal het beste past. Dus om een woord als *kast* te herkennen, moet de spraakherkenner akoestische modellen hebben van de /k/, de /ɑ/, de /s/ en de /t/ en moet het spraaksignaal het beste passen op deze modellen.

Taalmodel. Een taalmodel bevat informatie over welke woorden in welke volgordes voor kunnen komen in een gegeven taal. Het bevat informatie over hoe vaak een gegeven woord in een taal voorkomt (zo komen woordjes als *de* en *het* heel veel vaker voor in gesprekken dan woorden als *hoedenplank* of *deurklink*) en hoe vaak woorden in een bepaalde volgorde voorkomen. Uiteindelijk wordt die reeks van (in het lexicon aanwezige) woorden herkend, die het beste past op het spraaksignaal gegeven het akoestische spraaksignaal en de informatie in het taalmodel.

Hoe goed is een automatische spraakherkenner?

Hoe goed een automatische spraakherkenner is in het herkennen van spraak wordt standaard weergegeven in termen van *word error rate*, wat het percentage fouten is in een tekst, bijvoorbeeld een zin. Een spraakherkenner kan in zijn algemeenheid drie typen fouten maken:

- Inserties: dit zijn extra woorden die de spraakherkenner herkent die niet in de originele zin stonden. De spraakherkenner herkent dus meer woorden dan er uitgesproken zijn. Het gebeurt regelmatig dat een langer woord herkend wordt als twee kortere woorden.
- Deleties: dit zijn woorden die de spraakherkenner niet herkent. De spraakherkenner herkent dus minder woorden dan er uitgesproken zijn.
- Substituties: de spraakherkenner herkent een ander woord dan er uitgesproken is. De herkenner herkent evenveel woorden als er uitgesproken zijn.

Uiteraard kunnen er meerdere typen fouten in één zin voorkomen. Een voorbeeld:
Gesproken: de roos **is** een mooie bloem
Herkend: de *boos* een mooie een bloem

Is is een deletie, *boos* een substitutie en een een insertie. De *word error rate* voor deze zin is dan:

1 deletie + 1 substitutie + 1 insertie / 6 woorden in de gesproken zin = 3 / 6 = 50%

Een *word error rate* van 50% is overigens behoorlijk hoog. Voordat automatische spraakherkenningssoftware gebruikt kan worden zal de *word error*

rate (flink) moeten verbeteren. Dit is de afgelopen decennia gelukkig ook gebeurd.

In de afgelopen 65 jaar is het type spraak en de soort luisteromstandigheden waarvoor spraakherkenningssystemen gebouwd worden, flink veranderd. Begin jaren '90 hadden spraakherkenners een word error rate van bijna 100% voor spontane conversaties. Tien jaar later was dit gezakt naar ongeveer 30%. In 2016 werd een word error rate van 5,9% gehaald op dit type spraak door een systeem ontwikkeld door Microsoft. Het bijzondere van dit resultaat is dat deze onderzoekers ook menselijke luisteraars hebben laten luisteren naar de spraak die de herkenner moest herkennen (niet alle data, dit zou jaren kosten omdat het een vrij grote dataset is) en de mensen halen scores die *net* iets slechter zijn dan die gehaald door de automatische spraakherkenner! Dus, voor deze specifieke taak heeft de computer de mens bijgehaald!

Het is trouwens niet zo dat computers even goed als mensen zijn in het herkennen van spraak voor alle typen spraak, groottes van woordenlijsten en alle luisteromstandigheden. Over het algemeen genomen herkennen mensen de woorden in spraak nog steeds veel beter dan computers, en dit is vooral zo als er achtergrondgeluid aanwezig is in het spraaksignaal. Mensen doen het dan wel zo'n 6x beter dan automatische spraakherkenners.

Het genereren van spraak door computers: Spraaksynthese

Het doel van spraaksynthese is het automatisch omzetten van tekst naar gesproken taal, waarbij de tekst uitgesproken wordt met de juiste intonatie en met een goed klinkende stem. Dus als we een zin als *Emma wil wel warme chocomel* hebben, dan willen we dat de spraaksynthese iets oplevert zoals in het rechterpaneel van figuur 4. Spraaksynthese kan op verschillende manieren gedaan worden, maar de meest gangbare manier is het aan elkaar plakken van kleine stukjes spraak. Je kunt woorden aan elkaar plakken, stukken zin, hele zinnen, maar meestal worden er zogenaamde 'difonen' aan elkaar geplakt. Een voorbeeld van wat een difoon is: stel je hebt het woord *boom*. Dit woord bestaat uit drie klanken /b/, /o/, en /m/. Dit woord bestaat uit vier difonen, namelijk: <woordbegin>+/b/, /b/+/o/, /o/+/m/, /m/+<woordeinde>, waarbij één difoon steeds bestaat uit de achterste helft van de eerste klank en de voorste helft van de tweede klank. Het woord wordt dus niet geknipt *tussen* de klanken, maar in het *midden* van een klank.

Spraaksynthese via concatenatie van difonen bestaat ruwweg uit vijf verschillende stappen. De eerste stap is de tekstnormalisatie. In deze stap wordt de tekst omgezet naar een vorm die uitgesproken kan worden. Bijvoorbeeld een zin als:

De prijs van 1 t-shirt is EUR 3,20.

wordt omgezet naar:

de prijs van een1 t-shirt is drie euro twintig

waarbij leestekens en hoofdletters worden weggehaald, woorden worden gedisambigueerd (*een* kan zowel het lidwoord *een* zijn als het telwoord *één*) en getallen en afkortingen worden uitgeschreven. In de tweede stap worden de letters naar klanken omgezet. De bovenstaande zin wordt dan:

/dəprɛɪsvɑnentiʃərtɪsdriørotwɪntəx/

In de derde stap worden de juiste difonen uit een grote database gepakt en achter elkaar gezet. Deze difonen-database bevat in principe alle difonen van een taal en deze zijn uitgesproken door één spreker. In de vierde stap wordt de intonatie van de zin goed gezet: De klemtonen worden op de juiste plaats gelegd (vergelijk een uitspraak van bovenstaande zin als de klemtoon op *t-shirt* of op *drie* ligt) en hier kun je bepalen of de zin een stellende of vragende zin is (in het laatste geval gaat de intonatie op het einde van de zin iets om hoog). Tot slot wordt de golfvorm gegenereerd die hoort bij de opeenvolging van de difonen met de gemaakte intonatie.

Conclusie

Er zijn steeds meer spraakgestuurde toepassingen die het mogelijk maken om op een natuurlijke(re) manier met een machine of een robot te communiceren. In deze systemen wordt eerst herkend welke woorden er gezegd zijn door de gebruiker. Daarna

wordt de bedoeling van de gebruiker bepaald en wordt een actie ondernomen door de machine of robot.

Tot slot wordt er vaak weer feedback teruggegeven aan de gebruiker. Deze feedback kan ook in de vorm van spraak zijn die door een computer is gegenereerd. Dialogen met machines en robots via spraak zijn dus al mogelijk! Al kunnen hier nog wel wat fouten in optreden, met name als er veel geluid in de achtergrond aanwezig is.

Suggesties voor meer informatie

1. Jurafsky, D., & Martin, J.H., 2006, *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd edition)*. Chapter 8: Speech Synthesis, Chapter 9: Automatic Speech Recognition.
2. Van Oostendorp, M.: http://www.meertens.knaw.nl/medewerkers/marc.van.oostendorp/propedeuse/3.spraakorganen.html.

- - -