Spraaktechnologie: in de toekomst was alles beter

INAUGURELE REDE DOOR PROF. DR. IR. DAVID A. VAN LEEUWEN

Radboud Universiteit Nijmegen

INAUGURELE REDE

PROF. DR. IR. DAVID A. VAN LEEUWEN

Een belangrijk aspect van de spraaktechnologie is het automatisch ontrafelen van informatie uit het gesproken woord. Het streven is om automatisch te weten te komen wie wat zegt, in welke taal, op welke manier en met welke bedoeling.
Een mooi voorbeeld is sprekerherkenning ('Is dit bandje ingesproken door Bin Laden?'). We willen iemand kunnen herkennen, ongeacht wat hij of zij zegt, in welke taal, of het nu op *Al Jazeera* is of via de satelliettelefoon, en of de spreker in kwestie nu verkouden is of niet. Door telkens tests samen te stellen met grote hoeveelheden nieuwe sprekers, opgenomen in verschillende omstandigheden, worden wetenschappers uitgedaagd zo goed mogelijk presterende systemen te bouwen.
Spraaktechnologie staat midden tussen de wetenschap, waar nieuwe modelleringstechnieken en algoritmen moeten worden gevonden, en de toepassing, waar de spraakdata vandaan moeten komen. Zoals in het verleden al is gebleken, heeft het vak een mooie toekomst.

David van Leeuwen is sinds 1 mei 2008 bijzondere hoogleraar Spraaktechnologie en haar toepassingen aan de Radboud Universiteit. Van Leeuwen studeerde natuurkunde aan de Technische Universiteit Delft, promoveerde in Leiden en werkt sinds 1994 werkt als onderzoeker bij TNO Human Factors aan verschillende aspecten van spraaktechnologie. Als hoogleraar gaat hij zich bezighouden met het automatisch onttrekken van informatie uit spraaksignalen, in het bijzonder wat er gezegd wordt, wie er spreekt en in welke taal.

Radboud Universiteit Nijmegen

spraaktechnologie: in de toekomst was alles beter

Spraaktechnologie: in de toekomst was alles beter

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Spraaktechnologie en haar toepassingen aan de Faculteit der Letteren van de Radboud Universiteit Nijmegen op woensdag 19 november 2008

door prof. dr. ir. David A. van Leeuwen

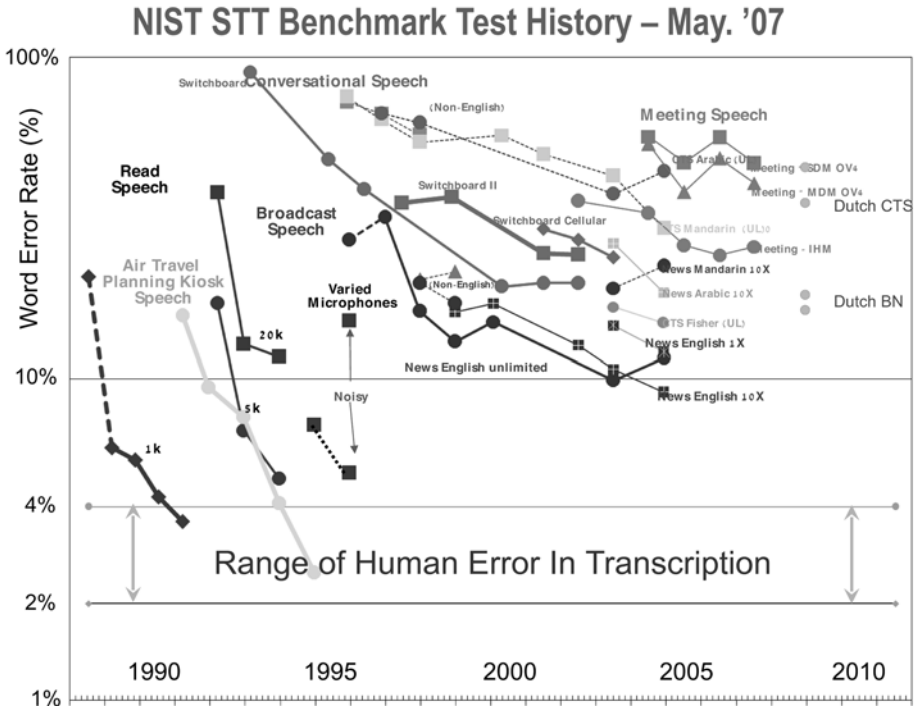
Vormgeving en opmaak: Nies en Partners bno, Nijmegen
Fotografie omslag: Bert Beelen
Drukwerk: Thieme MediaCenter Nijmegen

ISBN 978-90-9023756-5

Mijnheer de rector magnificus, zeer gewaardeerde toehoorders,

AL MAAR BETER

'Vroeger was alles beter.' Deze zin heeft u vast wel eens eerder gehoord. Meestal in een wat nostalgische setting, wellicht verzuchtend uitgesproken door een niet al te jeugdig persoon. Het zinnetje wordt zo vaak gebruikt, dat het wellicht als een Nederlandse uitdrukking kan worden opgevat. Het zinnetje suggereert een bepaalde context waarin het gebruikt wordt, en als zo'n situatie zich voordoet, ligt het gebruik van dat zinnetje ook voor de hand. Dit lijkt allemaal volstrekt triviaal, en dat is het waarschijnlijk ook, zij het dat deze overwegingen moeilijk zijn te gebruiken en te modelleren in een machine, bijvoorbeeld een apparaat dat gesproken woorden omzet in tekst. Zo'n apparaat noemen we een automatische spraakherkenner en het is de afgelopen decennia de heilige graal geweest van het onderzoek en de ontwikkeling in de spraaktechnologie om een spraakherkenner te maken die ingezet kan worden in de meest uiteenlopende omstandigheden



Figuur 1: Afnemende foutenpercentages in de loop van de tijd voor automatische spraakherkenning. Als de foutenpercentages te laag worden, wordt een moeilijker taak gesteld. Grafiek van Jon Fiscus, NIST.

en dan telkens goed presteert, zeg maar minstens zo goed als een mens dat doet. Ondanks het feit dat die heilige graal nog niet gevonden is, is het wel aangetoond dat de openingszin van deze rede niet geldt voor de spraaktechnologie zelf: in alle technologieën in de spraak waaraan gewerkt wordt worden in de loop van de tijd steeds betere prestaties bereikt. Vroeger was alles slechter – althans, in de spraaktechnologie [2, 5]

'Vroeger was alles beter'– als een spraakherkenner die getraind is voor het Nederlands dit te horen krijgt, is er toch een heel behoorlijke kans dat het correct herkend wordt. Dit komt, omdat deze combinatie van vier woorden zo vaak gebruikt wordt, dat ze ook regelmatig zo in de krant verschijnen. En krantenteksten vormen het traditionele trainingsmateriaal van het taalmodel van een spraakherkenner, het deel dat de *a priori* kansen van de woorden modelleert. *A priori* betekent dat een herkenner – zelfs zonder nog maar iets gehoord te hebben – al een verwachting heeft van wat er gezegd zal worden. Dat werkt net als bij mensen. U kunt dat bij uzelf uitproberen, door het volgende spelletje te spelen. Een bij uitstek geschikte situatie om dit te spelen is als u moet luisteren naar een lange rede, die u verder niet buitengewoon boeit. Probeer dan, terwijl u luistert, telkens te voorspellen wat het volgende woord is dat de spreker zal... zeggen. U zult merken, dat dat soms heel gemakkelijk... is, vooral aan het einde van de... zin. Wat daar aan de hand is in uw brein, is dat uw taalmodel u helpt bij het herkennen van de woorden die gesproken gaan worden. En in een spraakherkenner werkt het net zo.

AKOESTISCHE MODELLEN

'In de toekomst was alles beter.' Bij het horen van de titel van deze rede zal ook een spraakherkenner dus even verbaasd zijn: het woordje 'was' na de woorden 'in de toekomst' is volgens het taalmodel onverwacht.[1] Hier zal het systeem het dus vooral moeten hebben van het akoestische model. Dit is het deel van de herkenner dat de koppeling legt tussen woorden en hoe deze klinken. Over het algemeen heeft een herkenner modellen voor alle klanken – *foons* in het jargon – die voorkomen in een taal, en probeert door aaneenschakeling van klanken – de foons /w/, /a/ en /s/ vormen 'was'– het geluid te passen op de modellensequentie. Nu is er bij de uitspraak van woorden iets bijzonders aan de hand. Doordat onze tong, mond en lippen allerlei toeren moeten uithalen om de spraakklanken voldoende snel achter elkaar uit te spreken, beïnvloeden de foons elkaar in hun klank. We noemen dit coarticulatie. Een mooi voorbeeld is wellicht het oernederlandse woord 'washandje.' Dit woord bevat tweemaal de foon /a/, maar de realisaties – hoe ze klinken – verschillen enorm. Bij de tweede /a/ in 'handje' is de tong al op weg om de /j/ klank te realiseren – die pas drie foons na de /a/ aan de beurt is. Dit verandert de klank van de /a/ aanzienlijk, en om hem goed te kunnen herkennen zou eigenlijk een apart model gebruikt moeten worden. Dit is inderdaad wat een moderne herkenner doet: in plaats van één model voor de /a/ te gebruiken, wordt een apart model voor elke mogelijke context gemaakt. Als de context bestaat uit een klank voor en na de /a/ – we noemen dit een trifoon – dan zijn er al heel wat modellen nodig voor die foon. Als we ons realiseren dat het

Nederlands ongeveer 40 foons kent, komen we al op 1600 contexten voor de /a/ alleen. Om alle foons in het Nederlands op die manier te modelleren, zouden we weer 40 maal zoveel modellen nodig hebben, zo'n 64.000. Bij een trifoonmodel laat een foon zich alleen door zijn directe buren beïnvloeden, maar voor ons 'washandje' moesten we drie foons vooruit kijken, hiervoor zouden een astronomisch aantal heptafoonmodellen nodig zijn.

In de praktijk worden trifoonmodellen gebruikt, soms aangevuld met quinfoons, om de coarticulatie-effecten te kunnen modelleren. Het aantal modellen voor klanken wordt beperkt gehouden door slim stukken van de verschillende modellen gemeenschappelijk te maken. Toch kunt u zich voorstellen dat er op deze manier heel wat voorbeelden van de foons-in-context nodig zijn om een goed model te kunnen trainen. Een typisch akoestisch model wordt met honderden tot duizenden uren spraakmateriaal getraind, en van al dit materiaal is de letterlijke transcriptie nodig, die handmatig gemaakt moet worden. Voor het Nederlands is het *Corpus Gesproken Nederlands*, dat mede door inspanning van de Radboud Universiteit is gerealiseerd, een bruikbare bron hiervoor.

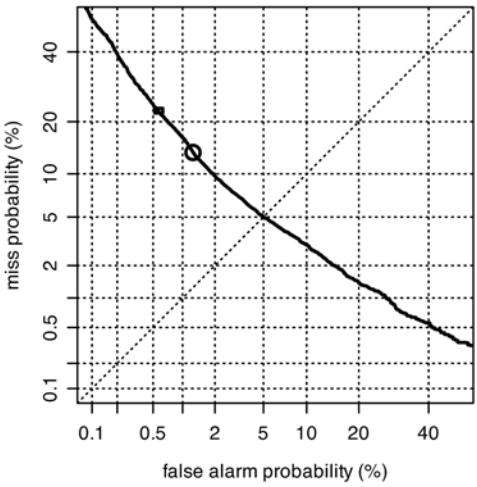
Behalve de al eerder genoemde coarticulatie-effecten zijn het de grote variaties in spraak die herkenning ervan lastig maken. Er is variatie in productie: spreker, spreekstijl, inspanning, accent, jargon, emotionele en fysieke gesteldheid van de spreker; er is variatie in akoestische omstandigheden: galm, achtergrondlawaai, en variatie ten gevolge van de opname: microfoon, transmissiekanaal en codering. Maar zoals we in Nederland weten heeft ieder nadeel weer z'n voordeel. We kunnen de variatie dus ook uitbuiten en onszelf een nieuwe taak toedichten: het herkennen van bijvoorbeeld spreker, accent of emotie uit het spraaksignaal.

SPREKERHERKENING

Ik wil u nu wat vertellen over *sprekerherkenning*, omdat het wellicht het mooiste probleem in de spraak is.[2] In de sprekerherkenning is het de taak om te bepalen of een spraaksegment is uitgesproken door een bepaalde spreker, of niet. Dit is een zeer eenvoudige vraag. En het antwoord is ja of nee. En dat antwoord is dan goed of fout. Wat willen we nog meer? De waarheid gebiedt me te zeggen dat het toch niet zo simpel is als het lijkt. Stel u voor dat ik een een fragment laat horen van de voormalig president Bill Clinton en daarbij de vraag stel: is dit Bill Clinton? De meesten van u zullen waarschijnlijk hierop positief antwoorden. Maar stel dan dat we hierover een weddenschap zouden afsluiten, en wel in de traditie van de Engelse *bookmakers* ongelijk zouden uitbetalen bij een correct antwoord: als het toch niet Clinton blijkt te zijn ontvangt u tien maal de inzet, maar als het inderdaad Clinton blijkt te zijn die sprak dan ontvangt u slechts tien procent bovenop uw inzet terug. Een deel van u zal dan waarschijnlijk het 'ja'-kamp verlaten, al was het maar voor de spanning.

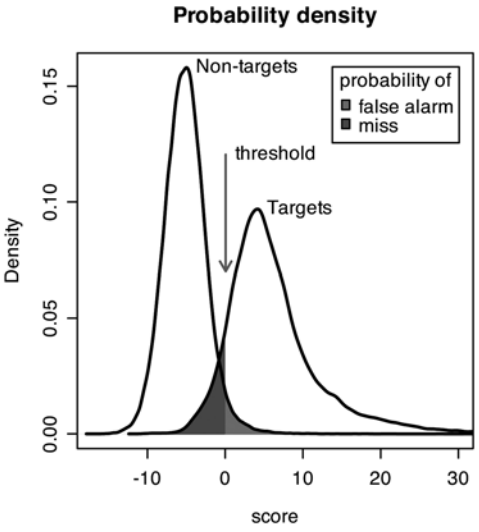
We zien dus dat er twee gevallen zijn bij een sprekerdetectievraag, namelijk wanneer het in werkelijkheid de gevraagde spreker is (we noemen dit een *target trial*), en wanneer dat niet zo is (een *non-target trial*). En hiervan moeten de antwoorden verschillend

behandeld worden. En de bonussen bij een goed antwoord – of equivalent, de kosten bij een fout antwoord – hebben een invloed op de keuze van het antwoord. Een sprekerherkenner werkt intern met een score die aangeeft hoe goed het spraakfragment bij de spreker past, en een beslissing wordt dan genomen door te kijken of een bepaalde drempelwaarde overschreden wordt. Door de drempel hoger te kiezen, verlagen we de kans op een vals positief, dus dat we ten onrechte zeggen dat het de spreker is, maar verhogen we de kans op een vals negatief, dat we ten onrechte beweren dat het niet de gevraagde spreker is. Deze afweging is mooi grafisch weer te geven in een zogenoemde DET grafiek [4] – *Dectection Error Trade-off* ofwel detectiefout afwegingscurve – zoals u hier ziet. Langs de assen ziet u de kans op een vals positief en vals negatief, respectievelijk, en de curve geeft de relatie tussen de twee weer wanneer de drempel van laag (rechtsonder in de grafiek) naar hoog (linksboven) verandert. Deze foutkansen zijn in dit geval bepaald door een sprekerherkenner – hier die van TNO – zo'n 100.000 maal een detectievraag voor te leggen, waarbij in het totaal zo'n 1300 sprekers voorkomen. Onderzoekers in de sprekerherkenning houden ontzettend veel van zo'n DET-curve. Je kunt namelijk snel zien of je het beter doet dan een ander door te kijken of jouw curve lager ligt dan die van haar – en of je dus een beter systeem hebt.



Figuur 2: Een DET-grafiek. De lijn geeft aan hoe de twee typen fouten, vals negatieven en vals positieven, tegen elkaar afwegen.

Een mooie rechte DET-curve geeft een onderzoeker het gevoel dat dingen kloppen. Een kromme DET-curve suggereert dat er iets vreemds aan de hand is. Inderdaad zijn de afgelopen jaren foutjes ontdekt in de veronderstelde identiteit van sprekers, die door sprekerherkenners aan het licht zijn gebracht vanwege kromme DET-curves. Fouten in de identiteiten kunnen bijvoorbeeld ontstaan doordat een spreker zich twee keer onder een verschillende naam inschrijft bij de datacollectie om zo meer te kunnen verdienen. Als de sprekerherkennerscores voor *target trials* normaal verdeeld zijn, en de scores van *non-target trials* ook, dan is een DET-curve recht. Dat komt, omdat de assen van een DET-curve zijn geschaald volgens de inverse cumulatieve normaalverdeling.[3]



Figuur 3: De distributies van scores voor *target trials* en *non-target trials*, die de DET-grafiek in figuur 2 oplevert.

Met de DET-curve kun je ook bepalen hoe je de drempel moet instellen al naar gelang je toepassing, en zien wat voor een prestaties je daar kunt verwachten. In een toepassing waar je op zoek bent naar de spreekwoordelijke naald in de hooiberg: één bepaalde spreker in een zee van audiofragmenten van anderen, zul je de drempel hoger moeten kiezen om je kostbare tijd niet te verdoen aan het napluizen van de vals positieven. Als je daarentegen op zoek ben naar fraude bij bijvoorbeeld telefonische transacties door de identiteit van de spreker te checken, zul je de drempel lager moeten kiezen, omdat de meeste gesprekken zullen worden gevoerd door de geautoriseerde persoon. Het feit dat de drempel afhangt van de toepassing, geeft al aan dat er een interactie is tussen toepassing en technologie.

Zoals u zult begrijpen werken we in de sprekerherkenningstechnologie niet met weddenschappen zoals in het voorbeeld van het fragment van Clinton. We werken met een zogenaamde kostenfunctie, die de fouten weegt met verwachte – abstracte – kosten. Het is de taak van de sprekerherkenner om de totaal verwachte kosten bij operationeel gebruik te minimaliseren. Dat kan natuurlijk het beste door de DET-curve zo laag mogelijk te hebben, dus door een goed onderscheid in scores tussen *targets* en *non-targets* te produceren. Maar ook van belang is het juist kunnen bepalen van de drempel – we noemen dit calibratie – anders worden er te veel fouten van één soort gemaakt waardoor de gemiddelde kosten hoger zijn dan nodig. Calibratie van een sprekerherkenner kan lastig zijn, vooral wanneer we te maken hebben met onverwachte situaties.

Laat ik dit proberen uit te leggen aan de hand van een vereenvoudigde analogie. Stel dat we onderscheid willen maken tussen voetgangers en fietsers, en het enige dat we van ze weten is hoe snel ze bewegen. We kunnen dan de snelheid van een groep fietsers en voetgangers meten, en een verstandige drempel kiezen die de kans op een fout minimaliseert. De enkele jogger en de moegestreden fietser daargelaten kan een drempel van 10 kilometer per uur heel aardig de twee groepen scheiden op basis van alleen hun snelheid. Nu gaat het systeem worden ingezet, maar op een onverwachte plek: vlak voor het bejaardentehuis. Ondanks het feit dat snelheid nog steeds een goed onderscheid kan maken tussen een lopende en de enkele fietsende bejaarde, is de drempel van 10 kilometer per uur veel te hoog waardoor bijna iedereen als voetganger wordt geclassificeerd.

CALIBRATIE

We zien dus dat calibratie belangrijk is, maar ook afhangt van de toepassing. Dat is onbevredigend, en daarom is enkele jaren geleden een nieuwe presentatie van het antwoord van een sprekerherkenner bedacht door de Zuid-Afrikaan Niko Brümmer [1]. In plaats van 'ja' of 'nee' als antwoord op de vraag: 'is dit spreker *X*' moeten we een *aannemelijkheidsverhouding* geven. Dit is een getal dat vergelijkbaar is met de eerdergenoemde interne score van de sprekerherkenner, maar het heeft een absolute interpretatie. Een aannemelijksverhouding van 10 betekent de kansverhouding (Engels: *odds*) voor spreker *X* 10 maal zo groot wordt door de spraak in aanmerking te nemen. Tien maal zo groot als wat dan? Wel, tien maal zo groot als de *a priori* kansverhouding dat het hier om *X* gaat, en niet iemand anders.

Ik heb het nog niet over de *a priori* kans gehad in het kader van *spreker*herkenning, maar hij vervult dezelfde rol als de *a priori* kansen op woorden voor een *spraak*herkenner. Waar in de spraakherkenning het modelleren van deze kans normaal is, we doen dat met het taalmodel, hebben we deze in het geheel weggewerkt in de sprekerherkenning. De *a priori* kans op een bepaalde spreker wordt als een gegeven beschouwd, en niet als een taak voor de herkenner om deze te schatten. Voor de *a priori* kans moeten we de spreker echt kennen, en weten hoe vaak zij belt, of op de radio is, of hoe we dan ook aan haar stemgeluid zijn gekomen – en dat ligt niet in het domein van de spraaktechnologie. Het is dus essentieel dat we met alleen een spraakfragment *nooit* antwoord kunnen geven op de vraag: 'wat is de kans dat hier spreker *X* spreekt,' omdat we de *a priori* kans niet kennen. We kunnen alleen een bestaande, extern toegekende kans groter of kleiner maken – de factor is de aannemelijkheidsverhouding – door gebruik van een goed gecalibreerde sprekerherkenner. Dit inzicht betekent dat gecalibreerde sprekerherkenning in principe ingezet kan worden als forensisch instrument, niet alleen in de opsporing, maar zelfs in de bewijsvoering. Op een vergelijkbare manier kan ook bewijsmateriaal van vingerafdrukken, gezichtsherkenning en DNA-sporen worden geïncorporeerd in dezelfde interpretatie. Merk hier overigens op, dat dit een zogenaamde Bayesiaanse interpretatie van de kansrekening veronderstelt.

Terugkomend op de aannemelijksverhouding, kunnen we nu ook wel meten of deze juist is geschat? Bij een harde beslissing ja of nee kunnen we deze als goed of fout rekenen, en met behulp van de kostenfunctie de merites van een sprekerherkenner beoordelen. Maar het blijkt dat er voor een 'zachte' uitkomst als de aannemelijkheidsverhouding, een logaritmische scoringsregel bestaat die als het ware de 'zachte' versie van de kostenfunctie is. Het verschil is dat de oude kostenfunctie een herkenner evalueert op één punt van opereren op de DET-curve, terwijl de nieuwe scoringsfunctie – we noemen hem C_{llr} – de herkenner over de gehele DET-curve evalueert. En net als dat er een minimale waarde voor de kostenfunctie bestaat – de kosten als de klassieke drempelwaarde ideaal gekozen was geweest – bestaat er een 'minimum C_{llr}' die aangeeft wat de C_{llr} had kunnen zijn als de calibratie over het gehele sprectrum van aannemelijkheidswaarden goed was geweest.

De nieuwe applicatie-onafhankelijke evaluatiemaat C_{llr} lost het calibratieprobleem niet op – maar we hebben in elke geval wel een middel om de kwaliteit ervan te meten. Voor een goede calibratie moeten we nog steeds werken met een oefenevaluatie, die ons kan vertellen wat een goede aannemelijkheidsverhouding had moeten zijn voor elke detectievraag. Hierbij vormen extreme hoge scores interessante gevallen. Als we boven een bepaalde score, zeg 10, alleen maar *target trials* hebben gevonden – laten we zeggen dat dit er 100 zijn – wat vertelt ons dit dan over de zekerheid van detectievragen waarbij de score groter is dan 10? Ben je dan honderd procent zeker? Deze vraag is vergelijkbaar met de vraag: als je de zon al honderd maal iedere morgen heb op zien komen, wat is de kans dat hij morgen opkomt? De Franse wiskundige Laplace heeft zich al eens over dit probleem gebogen, en na ingewikkelde integralen over bètadistributies is het antwoord vrij simpel: 101 tegen 1.

De nieuwe evaluatiemaat C_{llr} voor de presentatie van aannemelijkheidsverhoudingen en aanverwante algoritmes zijn alle door Niko Brümmer bedacht, en zijn ook geaccepteerd door het Amerikaanse NIST, National Institute of Standards and Technology, voor het evalueren van sprekerherkenningssystemen. Toch is het moeilijk om de internationale sprekerherkenninggemeenschap de merites ervan duidelijk te maken, en ik beschouw het dan ook als mijn missie om deze gemeenschap hiervan te overtuigen [6].

EVALUATIES EN DATA

Wat zijn nu de uitdagingingen in de automatische sprekerherkenning? Net als bij andere spraaktechnologieën gaat het weer om variatie. Bij tekstonafhankelijke sprekerherkenning verzorgen de woorden die gesproken worden in zekere zin een onbekende verstoring bovenop de karakteristieken van de stem die we juist willen modelleren en herkennen. Een andere verstoring is het transmissiekanaal. Als je iemand alleen nog maar in levende lijve hebt gesproken is het moeilijk om hem of haar te herkennen wanneer deze persoon je opbelt. Ik heb zelf ervaringen gehad waarbij de context me in zo'n geval zo op het verkeerde been zette – hier zie je hoe grote rol de *a priori* kans weer speelt – dat zelfs na het horen van de voornaam geen idee had wie ik aan de lijn had – iemand die ik enkele malen per week ontmoette. Het feit dat een ander kanaal – een telefoon, een codering (VoIP, GSM) – wordt gebruikt bij training dan bij herkenning heeft het afgelopen decennium het onderzoek gedomineerd. Hierop is dan ook in sinds 2004 enorme vooruitgang geboekt. Enerzijds door het ontwikkelen van nieuwe technieken, met het Factor Analyse model van Patrick Kenny als meest succesvolle maar ook meest ingewikkelde aanpak [3]. Anderzijds ook door het beschikbaar komen van training- en testdata waarbij dit kanaaleffect overheersend aanwezig is.

Hoe komen we aan al die data? Al in de jaren negentig van de vorige eeuw zag men in dat al die variabiliteiten in spraak alleen succesvol kunnen worden aangepakt door het opnemen en beschikbaar stellen van grote hoeveelheden spraakmateriaal. In de VS gebeurt dit door het *Linguistic Data Consortium*, in Europa is later de *European Language*

Resources Association met dit doel opgericht. Deze data wordt vaak opgenomen in het kader van een specifieke technologie-evaluatie, die in de vs georganiseerd worden door het al eerder genoemde NIST. In de sprekerherkenning zijn sinds 1996 vrijwel jaarlijks evaluaties georganiseerd waarin deelnemende systemen telkens nieuwe data voorgeschoteld krijgen – nieuwe sprekers, nieuwe condities – die zij dan moeten verwerken volgens een protocol. Herkenneruitvoer wordt opgestuurd naar NIST, die scoort deze, en de resultaten worden dan besproken in een *workshop*. Onderzoekers met de beste prestaties mogen het eerste presenteren – alle anderen willen natuurlijk weten wat ze hadden moeten doen om daar te staan. Deze enigszins competitieve aanpak leidt ertoe dat onderzoekers enorm hard werken om nog voor de deadline zo goed mogelijke antwoorden in te leveren, om dan daarna met spanning te wachten op de resultaten. De grootste stappen voorwaarts in technologie vinden dan ook altijd vlak voor een evaluatie plaats.

Een bijkomstig effect van dit soort technologie-evaluaties is dat er almaar meer en nieuwe data komen. *Meer* is goed omdat dat beter modellen toelaat, *nieuw* is goed omdat er in het in het vakgebied van het machineleren het gevaarlijk is om met oude data te evalueren: dit kan tot overtraining, en te optimistische resultaten, leiden. Een belangrijke reden voor onderzoekers om mee te doen aan dit soort evaluaties, en een maand van je leven op te offeren, is dat je de nieuwe data en kennis van anderen veel eerder krijgt dan via reguliere kanalen. Omdat er zo veel partijen meedoen aan deze evaluaties, krijgen deze een bijzondere betekenis in de literatuur: als iemand over de NIST sprekerherkenningevaluatie van 2006 schrijft weten haar *peers* precies om wat voor data het gaat, en dit maakt het interpreteren van resultaten veel eenduidiger. De vs krijgen hun investeringen in data en infrastructuur vele malen terugbetaald in de vele maanden inspanning van de beste onderzoekers op de wereld die nieuwe technieken bedenken, implementeren, testen en publiceren.

Dit Amerikaanse model is zo succesvol dat op bescheiden wijze hier in Europa het voorbeeld wordt gevolgd. Noemenswaardig zijn onderzoeken waarbij ik bij TNO een rol heb mogen spelen als coördinator. In het EU project SQALE zijn spraakherkenningssystemen in vier talen geëvalueerd. In het STEVIN-project N-Best, waaraan de Radboud Universiteit op verschillende manieren heeft bijgedragen, zijn voor het eerst spraakherkenners in het Vlaams en Nederlands langs de lat gelegd. In de NFI-TNO forensische sprekerherkenningsevaluatie zijn een tiental systemen wereldwijd getest met spraak van echte boeven, verkregen uit telefoontaps van de politie. Doordat ik zelf ook vaak meegedaan heb met NIST-evaluaties, vaak in samenwerking met andere instituten, weet ik hoe stimulerend en leerzaam deze inspanningen zijn. Ik zal er daarom naar streven dat ook de Radboud Universiteit vaker mee zal doen in de evaluaties van de diverse spraaktechnologieën.

In dit kader kan ik ook noemen dat het van belang is om *state-of-the-art* systemen te gebruiken. Maar al te vaak lijkt een nieuwe techniek een kleine verbetering aan te

brengen in een systeem, terwijl het in werkelijkheid alleen maar een onvolkomenheid deels repareert. Voor een *state-of-the-art* systeem zonder deze onvolkomenheid blijkt zo'n nieuwe techniek dan ineens niets meer op te leveren.

TOEKOMSTIG ONDERZOEK

We weten dat een sprekerherkenner beter presteert naar mate er meer spraak beschikbaar is – in training of tijdens detectie. Dat ligt voor de hand, maar als fysicus zou ik graag willen begrijpen hoe die afhankelijkheid dan precies is. We weten bijvoorbeeld dat de verdeling van foons en woorden in de taal aan bepaalde machtswetten voldoen – dat zouden we moeten kunnen uitbuiten om een model te maken voor de duurafhankelijkheid van de herkennerprestaties.

Zo hebben we eens een schets gemaakt van een model voor hoe goed mensen zijn in het herkennen van een *taal*. We zijn hierin geïnteresseerd, omdat vergelijking van een technologie met menselijke prestaties inzichten kunnen geven over wat de mogelijke limieten van de technologie zijn, en omdat het iets kan zeggen over hoe mensen talen herkennen. Dit model is al discussiërend op een wandbord ontstaan, en heeft geleid tot een onderzoeksopzet die meet wat de menselijke taalherkenningsprestaties zijn naarmate men de taal in kwestie beter kent. De uitkomsten van dit onderzoek laten zien dat dit eenvoudige model helemaal niet onaardig was, al ontbreekt het hier wel aan een goede mathematische ondergrond voor het model.

Een andere onopgelost probleem uit de sprekerherkenning is de stijlheid van de DET-curve. We zien meestal dat systemen in de loop van de tijd vlakkere DET-curves laten zien – dat is niet gek, want in het evolutionaire ontwikkelproces worden vlakkere curves beloond omdat ze een lagere NIST-kostenfunctie hebben. Maar af en toe, en de NIST-sperkerherkenningsevaluatie 2006 is een mooi voorbeeld hiervan, komen de curves weer overeind, en dit zien we dan over de hele breedte van het veld. Wat is er anders geworden in de data die de DET-curves doen roteren? En kunnen we de stijlheid van de DET onder controle krijgen?

En zo zijn er meer vragen. Hoe werkt sprekerherkenning bij de mens? Hoe gaan we zinnig met heel grote of juist kleine *a priori* kansen om? Hoe goed kunnen we een stem van tien jaar terug herkennen? Hoe kunnen we de calibratie goed krijgen van één enkele detectietest?

SPRAAKTECHNOLOGIE EN HAAR TOEPASSINGEN

Ik heb tot nog toe slechts over enkele technologieën in de spraak gesproken. We kunnen niet alleen spraak en spreker herkennen, maar ook taal en andere paralinguistische zaken zoals accent en emotie. Bij accent- en emotieherkenning is het overigens moeilijk om de werkelijkheid te kennen, die van belang is voor de technologen. Tegenover herkenning staat synthese. Spraaksynthese is een onderzoeksgebied dat al vele toepassingen kent. De verstaanbaarheid is al geen probleem meer, er wordt meer gewerkt aan kwaliteit,

natuurlijkheid en expressie, en methoden om snel nieuwe stemmen te creëren. Een belangrijk vak is ook de spraak codering – dat is hoe een mobieltje spraak digitaal inpakt, en hoe MP3 muziek inpakt – waar de trend is coders te maken met lagere vertraging en bitsnelheid. Een codeersysteem heeft altijd een vertraging, onder andere omdat voor een goede analyse van het spraaksignaal een stukje vooruit in de tijd gekeken moet worden – maar omdat dat niet kan worden stukjes bewaard en wordt de toekomst achteraf gebruikt – in de toekomst was alles beter. Als laatste wil ik stemtransformatie noemen, die als doel heeft de stem van één spreker te leggen over de spraak van een ander. Met wortels in de spraakcodering en spraaksynthese, vormen dit soort technologieën ook weer interessante onderzoeksonderwerpen voor bijvoorbeeld sprekerherkenning.

Spraakherkenning kent ook al vele toepassingen. Gepersonaliseerde dicteersystemen zijn al een decennium goed bruikbaar. Een reden dat dicteersystemen niet zo veel gebruikt worden als andere kantoorapplicaties is dat mensen helemaal niet gewend zijn om te dicteren: het valt helemaal niet mee om een zin in een keer uit te spreken zoals je hem op papier wil hebben. Je ziet dan ook dat dicteersystemen vooral gebruikt worden bij beroepsgroepen die dicteren gewend zijn, zoals radiologen en juristen. Voor spraakherkenning waar de marge voor verbetering groter is, bijvoorbeeld bij sprekeronfhankelijke transcriptiesystemen, is een belangrijke toepassing de gesproken documentontsluiting. Sinds het begin van deze eeuw is talloze keren gedemonstreerd dat zelfs met systemen die de helft van de woorden onjuist herkennen, nog een zinvol zoeksysteem is te maken. Een positieve kijk op deze performance is namelijk dat de andere helft van de woorden goed is herkend, en daarmee kun je audiofragmenten vinden bij zoektermen, net zoals we informatie op het web kunnen googelen. In figuur 4 ziet u een voorbeeld van een zoekterm die in het nieuws was ten tijde van het schrijven van deze rede, en hoe de spraakherkenner radiofragmenten kan vinden die hierover gaan. Het is aardig om te vertellen dat nu na jaren ontwikkeling en enkele promotiestudies in Nederland, het Nederlands Instituut voor Beeld en Geluid deze technieken gaat gebruiken om hun audiovisuele archieven beter te ontsluiten.

Het systeem dat de spraak van het Radio1-Journaal heeft herkend, en waarvan u nu delen van de transcriptie ziet, is een zogenaamd recurrent neuraal net. Het koppelt de uitvoer van het net met enige vertraging weer terug in de invoer, en op die manier kan het de coarticulatiecontext in het verleden modelleren. Maar met het 'washandje' hebben we gezien dat we ook context de andere kant op moeten modelleren. Dus dit systeem draait het spraaksignaal ook om in de tijd, om een achterstevoren getraind neuraal net de akoestische context in de toekomst te laten modelleren. We kunnen een spraaksignaal pas achteraf omkeren om tot betere prestaties te komen – in de toekomst was alles beter.

Figuur 4: Een voorbeeld van een zoekopdracht naar het recent voorkomen van de woorden 'krediet' en 'crisis' in uitzendingen van het Radio1-Journaal

DANKWOORD

Aan het einde van deze rede wil ik enkele woorden van dank uitspreken. Allereerst gaat mijn dank uit naar het stichtingsbestuur en het college van bestuur van de Radboud Universiteit Nijmegen voor het in mij gestelde vertrouwen. Om eenzelfde reden bedank ik de commissieleden van de Stichting Lorentz-van Iterson Fonds TNO, die deze bijzondere leeropdracht heeft ingesteld. Zonder de inspanningen van Tini Colijn, Peter Werkhoven, Hans Godthelp, Adelbert Bronkhorst en Judith Kessens, maar ook Steve Renals en Dirk van Compernolle, had de commissie zich niet over mijn kandidatuur hoeven buigen.

Ik besef dat ik in de gelukkige omstandigheid verkeer dat ik in mijn professionele bezigheden bij TNO voldoende de kans heb gekregen om me te ontwikkelen als een spraakonderzoeker. Mijn eerste baas Herman Steeneken liet vanaf het begin blijken het volste vertrouwen in mij te hebben. Tevens heeft hij me geïntroduceerd bij vele spraakonderzoekers in Europa, waardoor ik vanaf het begin in deze carrière met de groten in dit onderzoeksgebied heb kunnen werken. Veel heb ik ook meegekregen van Sander van Wijngaarden, zonder meer de meest talentvolle en veelzijdige collega waar ik mee heb mogen samenwerken.

Hooggeleerde Boves, beste Lou. Eén van onze eerste gezamenlijke activiteiten was het schrijven van een Europees Handboek – ik had toen als nieuwkomer bij TNO al meteen het gevoel van wederzijds respect. Ik ben blij dat ik nu in jouw leerstoelgroep kan werken aan het verwerven van onderzoeksprojecten en het begeleiden daarvan, en ik denk dat ik op dit gebied in de komende jaren nog veel van je kan leren.

Hooggeleerde Bronkhorst, beste Adelbert. Jij bent als afdelingshoofd zonder meer stimulerend geweest in het bereiken van wetenschappelijke mijlpalen zoals deze dag er een voor mij is. Ik heb bij TNO kunnen genieten van werken aan grotere projecten, onder andere verworven door collega's Franciska de Jong en Wessel Kraaij, waardoor er de ruimte was om moeilijker problemen aan te pakken en betere technologie te ontwikkelen. Ik hoop in de toekomst met jullie ook weer samen te kunnen werken.

Hooggeleerde Van Leeuwen, beste Hans. In Nijmegen geboren, waar jij toen werkte aan de Universiteit, ben ik je in geografische zin gevolgd in Delft en Leiden, om nu weer hier terug te keren. Alice en jij hebben altijd gezorgd voor een stabiele omgeving waarin wij kinderen op konden groeien en jullie hebben altijd elke vorm van opleiding en ontwikkeling gestimuleerd – dat zijn slechts enkele van de dingen waar ik jullie heel erg dankbaar voor ben.

Zeergeleerde Orr, lieve Rosemary. Het is mij een eer om je in dezelfde zaal waarin jij je proefschrift hebt verdedigd te mogen bedanken voor de steun die jij me afgelopen tien jaren hebt gegeven. Zonder jouw onvoorwaardelijke positieve houding bij keuzes en plannen in het leven en daadkrachtig optreden op momenten dat het er toe doet zou ik niet zo ver gekomen zijn. Lieve Tess en Jan Fiach – jullie hebben de toekomst. Het is een genot te zien hoe jullie die invullen en beleven – en je weet het nu: in die toekomst was het allemaal beter.

Ik heb gezegd.

NOTEN

1 De herkenner staat als het ware perplex, we zeggen dat de perplexiteit hier omhoog gaat.

2 Zie [6] waarom het zo'n mooi probleem is.

3 Het omgekeerde is niet noodzakelijkerwijs het geval — een rechte curve impliceert niet dat scoredistributies normaal verdeeld zijn.

ENKELE REFERENTIES

[1] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20:230–275, 2006.

[2] Jonathan G. Fiscus, Nicolas Radde, John S. Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun. The Rich Transcription 2006 spring meeting recognition evaluation. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 309–322. Springer Berlin/Heidelberg, 2007.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(4), May 2007.

[4] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech 1997*, pages 1895–1898, Rhodes, Greece, 1997.

[5] David Pallett. A look at NIST's benchmark ASR tests: Past, present, and future. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 483–488, 2003.

[6] David A. van Leeuwen and Niko Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In Christian Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg - New York - Berlin, 2007.