Trillende stem, bloeiende leugen: over stemanalyse als vorm van leugendetectie

Harald Merckelbach en Ewout Meijer

Naar verluidt maken honderden Amerikaanse politiekorpsen er gebruik van. Ook verzekeringsmaatschappijen zetten het apparaat in. We hebben het over apparatuur die op basis van stemanalyse tot een oordeel komt over de waarachtigheid van verdachten of cliënten. De auteurs schaften zo'n apparaat aan en namen de proef op de som.

Eind 2003 belde misdaadjournalist Peter R. de Vries aan bij het huis waar Paul de Rijk zich op dat moment al een tijdlang ophield. Niet dat het zijn huis was: het behoorde toe aan het Baarnse echtpaar Muller, dat daar al jaren niet meer was gezien. De Rijk vertelde bezorgde buurtbewoners en familieleden dat hij op het huis van het echtpaar paste. De Mullers zouden zich in België hebben gevestigd. Om een oogje in het zeil te houden had De Rijk op hun verzoek zijn intrek in de Baarnse woning genomen. Althans, dat zei De Rijk. De politie geloofde hem. En waarom ook niet? De Rijk zag eruit als een vriendelijke oudere heer en hij maakte zich ook al verdienstelijk bij de plaatselijke kinderboerderij.

De Rijk liet de misdaadverslaggever binnen. In de woonkamer ontspon zich een interessant vraaggesprek, dat door Peter R. de Vries werd vastgelegd met een verborgen camera. De Rijk zei dat meneer Muller af en toe nog wel langskwam om de post op te halen en dan berichten achterliet, maar dat hij, De Rijk, geen telefoonnummer of adres van de familie had. Loog De Rijk?

Stem en leugendetectie

Sommige bedrijven beweren dat zo'n vraag zich bij uitstek laat beantwoorden met een speciale vorm van leugendetectie. Ze brengen apparatuur op de markt waarmee onderzoekers – politiemensen, journalisten, verzekeringsagenten, inlichtingenofficieren – het stemgeluid van hun gesprekpartners kunnen analyseren. Die analyse zou veilige conclusies toelaten over de waarachtigheid van de gesprekspartners. Dat is althans wat de fabrikanten van apparaten met klinkende namen als de *Computer Voice Stress Analyzer*, *Vericator* en *Truster* schrijven in hun wervende folders. Ze stuiten daarmee niet meteen op ongeloof. Geschat wordt dat in de VS enkele honderden politiekorpsen met enige regelmaat zulke apparaten inzetten. Meer omstreden is het gebruik van stemanalyse bij de ondervraging van vermeende terroristen in Irak en Guantanamo Bay. Het verhaal gaat dat een medewerker van het bedrijf dat de *Computer Voice Stress Analyzer* op de markt brengt, ene Bill Endler, de kans kreeg om er een voormalige Iraakse vicepresident mee aan de tand te voelen. De vice-president zou allerlei terroristische aanslagen op de coalitietroepen in Irak hebben opgebiecht nadat Endler hem had gedemonstreerd dat het apparaat perfect onderscheid kan maken tussen waarheid en leugen. Niettemin heeft het *Pentagon* leugendetectie via stemanalyse verboden omdat het wetenschappelijk gezien te prematuur zou zijn. Dat verbod wordt overigens door een aantal senatoren betwist.[1]

Waaraan stemanalyse haar populariteit te danken heeft, is ondertussen wel duidelijk. Bij de traditionele leugendetector wordt de verdachte via elektrodes aangesloten op de apparatuur. Dat is een tijdrovende procedure, die de medewerking van de verdachte vereist. Bovendien beseft de verdachte heel goed wat er gaat gebeuren en kan hij daarom zogenaamde *counter measures* nemen. Dat wil zeggen: de verdachte

Prof. dr. H.L.G.J. Merckelbach, psycholoog, Faculteit der Psychologie van de Universiteit Maastricht, Postbus 616, 6200 MD, Maastricht, h.merckelbach@psychology.unimaas.nl; *drs. E.H. Meijer*, rechtspsycholoog, Faculteit der Psychologie van de Universiteit Maastricht.

kan op zijn tong bijten of zijn voet in een kramptoestand brengen en daarmee de fysiologische signalen die worden gemeten, totaal ontregelen. Deze problemen zijn te omzeilen met stemanalyse, want die vorm van leugendetectie kan in principe plaatsvinden zonder dat de verdachte er weet van heeft (Merckelbach & Boelhouwer, 2002). Het is dus een discrete vorm van leugendetectie, die ideaal lijkt voor het screenen van passagiers op vliegvelden, het evalueren van schadeclaims die telefonisch worden aangemeld bij verzekeraars, het beoordelen van sollicitanten door werkgevers en het ondervragen van verdachten of overspelige partners. Maar dan de cruciale vraag: werkt het ook?

'Yes, I like this picture'

Het idee dat stemgeluid iets zegt over de waarachtigheid van de boodschapper is intuïtief plausibel. Iedereen kent immers de ervaring van de overslaande stem. Daarnaar verwijzend schreef de Latijnse aforist Publilius Syrus al dat de stem de spiegel van de ziel is. Daarvoor bestaat ook een biologische rationale. Die werd geformuleerd door de veelzijdige Britse fysioloog O.C.J. Lippold. Hij beschreef in de jaren zestig van de vorige eeuw het fenomeen dat armspieren in een ontspannen toestand spontane contracties met een frequentie van 8-12 per seconde vertonen. Onder stressvolle omstandigheden worden deze microtremoren minder prominent als gevolg van vaatvernauwing (Gamer et al., 2006). Het waren twee gepensioneerde inlichtingenofficieren van het Amerikaanse leger die vervolgens met het fenomeen aan de haal gingen. Zij redeneerden dat het ook moest gelden voor de spieren betrokken bij de menselijke stem en dat langs die weg leugendetectie mogelijk was.[2] Hun apparaat – de *Psychological Stress Evaluator* – registreerde de activiteit van de stem in de 8-12-frequentieband. Was die activiteit sterk, dan sprak de onderzochte persoon de waarheid; bleef die activiteit aan de magere kant, dan werd er gelogen. Aldus de ontwerpers van de *Psychological Stress Evaluator* (zie Lykken, 1998).

In zijn standaardwerk over leugendetectie heeft de Amerikaanse psychofysioloog David Lykken geen goed woord over voor de *Psychological Stress Evaluator*. In hoofdstuk 13 van *A Tremor in the Blood* maakt hij de pretenties van het apparaat met de grond gelijk. Daarbij verwijst hij onder andere naar een in opdracht van de Israëlische politie uitgevoerd experiment dat tamelijk fataal uitpakte voor de *Psychological Stress Evaluator*. Het was een simpel en elegant experiment dat hierop neer kwam: proefpersonen keken naar twee soorten dia's: mooie landschappen en gemutileerde lichamen. Elke dia bleef vijf seconden staan en daarna moesten de proefpersonen zeggen: 'Yes, I like this picture'. Terwijl ze die zin uitspraken, draaide de *Psychological Stress Evaluator* mee. Drie 'blinde' maar wel speciaal opgeleide experts bekeken de uitslagen van het apparaat en moesten op basis daarvan aangeven wanneer de proefpersonen logen. Geen van de experts kon meer dan 50% van de zinnen juist – waar of gelogen – classificeren. Daarnaast waren zij het in 60% van de zinnen oneens over de vraag of de waarheid werd gesproken. In de handen van deze experts kwam de *Psychological Stresss Evaluator* dus nauwelijks uit boven het niveau dat wordt gehaald als men de waarheid via het opgooien van een munt probeert te bepalen. Niet alleen dit Israëlische experiment kwam tot die conclusie. Ook andere studies pakten teleurstellend uit voor de *Psychological Stress Evaluator* (zie voor een kort overzicht: Krapohl et al., 2002).

Nieuwe versies

Sinds de jaren tachtig van de vorige eeuw verschenen nieuwe leugendetectieapparaten op basis van stemanalyse.[3] Kenmerkend voor deze jongste generatie van apparaten is dat zij zich niet beperken tot metingen van de hoeveelheid activiteit in de 8-12 Hz-band, maar ook andere frequentiedomeinen van de menselijke stem registreren en meenemen in hun analyse. De aanname is dat zo'n fijnmaziger analyse meer succes garandeert bij het ontmaskeren van leugenaars. En inderdaad, de pretenties zijn niet van de lucht. De website van één fabrikant meldt bijvoorbeeld: 'Ons apparaat is effectief in het opsporingsonderzoek naar moorden, zedenzaken, roofovervallen, witte boorden criminaliteit (...). Het systeem heeft bewezen een zeer betrouwbaar onderzoeksinstrument te zijn bij het verifiëren van verklaringen.' De handleiding van een andere fabrikant omschrijft haar apparaat als: 'an innovative, highly advanced, computerized system that is especially designed to provide you with easy access to truth verification.' Om na te gaan of zulke loftuitingen enige grond hebben, besloten we zelf de proef op de som te nemen. We kochten een hypermodern stemanalyseapparaat ter waarde van $10.000, bestudeerden zeer nauwgezet de handleiding, installeerden de bijgeleverde software[4] en zetten vervolgens een experiment op.[5]

Het experiment

Het stimulusmateriaal voor het experiment werd verschaft door 24 vrijwilligers (van wie 19 vrouwen) met een gemiddelde leeftijd van 23 jaar. Zij

kregen het verzoek een verhaal op papier te zetten over iets wat ze echt hadden meegemaakt en dat hen had aangegrepen. Het mocht gaan over een negatieve maar ook een positieve gebeurtenis. Het verhaal moest binnen het bestek van 400 woorden worden geschetst. De verhalen die we kregen beschreven acht positieve gebeurtenissen (bijvoorbeeld: ik won een reis naar Griekenland) en 16 negatieve gebeurtenissen (bijvoorbeeld: de dood van mijn grootvader). Vervolgens vertelde elke vrijwilliger uit de losse pols het eigen verhaal, maar ook – en dat op basis van de tekstuele versie die men van tevoren had geraadpleegd – een verhaal dat afkomstig was van een andere vrijwilliger. Elke vrijwilliger presenteerde zodoende een verhaal dat hij/zij wél had meegemaakt (waar verhaal) en een verhaal dat hij/zij níet had meegemaakt, maar wel als zodanig bracht (onwaar verhaal).Ware en onware verhalen werden op video en audio opgenomen, evenals een korte introductie waarin elke vrijwilliger zich voorstelde. Dat laatste diende ter calibratie van het apparaat. Natuurlijk balanceerden we de volgorde van de ware en onware verhalen.

De audio-opnames werden vervolgens ter verificatie aan het stemanalyseapparaat aangeboden. De vraag was of het apparaat goed onderscheid kon maken tussen ware en onware verhalen. Maar wat is een goed onderscheid? Het antwoord is helder: het apparaat zou het beter moeten doen dan mensen. Alleen dan heeft het zin om zo'n stemanalyseapparaat aan te schaffen en in te zetten. Daarom vroegen we ook aan 20 proefpersonen (10 vrouwen) met een gemiddelde leeftijd van 23 jaar om goed naar de videoopnamen van 24 verhalen (12 waar en 12 onwaar) te kijken en te luisteren. En ons na elk verhaal te vertellen of het waar of gelogen was. We verzochten hun tevens telkens op een 100-puntsschaal aan te geven hoe zeker ze van hun oordeel waren (0 = ben hier totaal onzeker over; 100 = ben hier absoluut zeker van). We verzamelden deze gegevens om zo een goede vergelijking met het stemanalyseapparaat mogelijk te maken.

Mensen zijn beter

De tabel laat de 2-x-2-matrix zien waarin werkelijkheid (waar en onwaar) en oordelen van proefpersonen (waar en onwaar) tegen elkaar worden uitgezet. De proefpersonen classificeren 63% van de ware videofragmenten en 47% van de onware fragmenten correct, wat hun diagnostische nauwkeurigheid op 55% brengt. Dat is iets beter dan kans (50%), waarbij de aantekening past dat de proefpersonen nauwkeuriger zijn in het detecteren van de waarheid dan in het detecteren van onwaarheden ($p = .02$). Deze asymmetrie staat in de literatuur bekend als de *truth bias* (DePaulo et al., 1997): mensen neigen ertoe aan te nemen dat anderen vaker de waarheid vertellen dan dat ze liegen.

Het stemanalyseapparaat beoordeelt videofragmenten niet in termen van waar of onwaar, maar geeft zogenaamde waarheidspercentages. Dit percentage duidt aan hoeveel procent van het gespreksfragment uit waarheid bestaat, en varieert dus van 0 (hele fragment onwaar) tot 100% (hele fragment waar). Om dit percentage met het oordeel van proefpersonen te kunnen vergelijken, transformeerden we de scores van proefpersonen zo dat ook deze een bereik van 0 (zeker gelogen) tot 100 (zeker waar) hadden.[6] Zoals de figuur op de volgende pagina toont, presteren de proefpersonen ook bij deze benadering boven kans. Zij schatten de ware videofragmenten iets hoger in dan de onware (59% vs 53%, $p = .04$). Hoe anders ligt dat voor het stemanalyseapparaat. Volgens het apparaat bestaan zowel de ware als de onware videofragmenten voor 95% uit waarheid en voor 5% uit onwaarheden (d.w.z. kleine onjuistheden). Hier is derhalve sprake van een uitgesproken *truth bias*. Dit brengt de diagnostische scherpte van het apparaat op kansniveau (50%). De moraal van het verhaal is dat het apparaat het slechter doet dan menselijke beoordelaars.

Is het dan allemaal flauwekul?

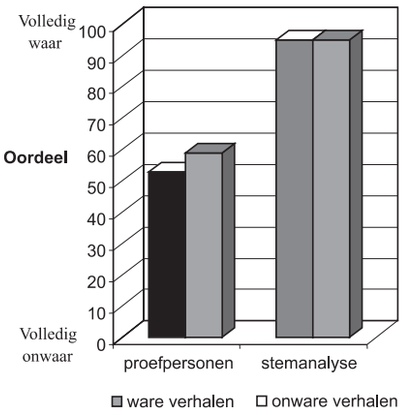
Net als onze voorgangers in het onderzoek (Krapohl et al., 2002; Gamer et al., 2006) vonden we in ons experiment geen enkel aanknopingspunt voor de hoge verwachtingen die fabrikanten met hun stemanalyseapparatuur bij de consument wekken. Het ding doet niet wat het belooft te doen, namelijk een goed onderscheid maken tussen waarheidsprekende en liegende mensen.

Waar komen de succesverhalen dan

Proefpersonen	oordelen 'waar'	oordelen 'onwaar'
Waar fragment	63%	37%
Onwaar fragment	53%	47%

Tabel Percentage ware en onware videofragmenten dat door proefpersonen correct (en foutief) wordt geclassificeerd.

vandaan waarmee de fabrikanten van de stemanalyseapparatuur op hun websites en in hun brochures goede sier maken? Wie de verhalen goed bekijkt, ziet wat er aan de hand is: de successen zijn niet gebaseerd op de diagnostische scherpte van de apparatuur maar op de intimiderende werking die ervan uitgaat. Net zoals het geval was bij de voormalige Iraakse vice-premier, gaan sommige ondervraagden door de knieën zodra ze door hebben dat ze met een leugendetector worden gescreend. Dat heet het *bogus-pipeline*-effect (Merckelbach & Jelicic, 2005). Die naam verwijst naar een leugendetectieapparaat dat bestaat uit een lege maar indrukwekkend ogende doos waarmee de verdachte via wat nutteloze kabels – de *bogus pipe lines* – is verbonden. Als zo'n verdachte bekent, is dat niet de verdienste van het apparaat maar van de wijze waarop het wordt gepresenteerd aan naïeve verdachten. Dat effect beogen ook verzekeringsmaatschappijen die het apparaat bij hun *call centers* inzetten: cliënten die hun schadebehandelaar bellen, krijgen dan te horen dat een stemanalyse wordt gemaakt. Op deze wijze zijn



Figuur Oordelen van proefpersonen en het stemanalyseapparaat (0-100%) over de waarachtigheid van de fragmenten

Engelse verzekeraars erin geslaagd het aantal valse schadeclaims behoorlijk terug te brengen.[7]

Betekent het allemaal dat de fysiologische rationale achter stemanalyse flauwekul is? Toch niet helemaal. Neem de meta-analyse van DePaulo en collega's (2003). Hun analyse omvat 120 studies waarin werd gekeken naar verbale en non-verbale verschillen tussen waarheidsprekende en liegende mensen. Leugenaars bleken inderdaad met een hogere stem te spreken dan eerlijke mensen. Maar de *effect size* van dit verschil was bescheiden (d = 0.21) en lag bijvoorbeeld ver onder dat voor pupildilatie (leugenaars hebben wijdere pupillen dan waarheidssprekende mensen; d = 0.39). Die matige *effect size* laat het hele probleem zien: het verschil in stemfrequentie tussen eerlijke en liegende mensen is weliswaar significant, maar te subtiel om op basis daarvan in individuele gevallen te oordelen over de waarachtigheid van de spreker.

Nogmaals *truth bias*

Toen misdaadverslaggever Peter R. de Vries De Rijk opzocht in het huis van het Baarnse echtpaar had hij behoorlijk wat voorwerk verricht. Zo wist hij bijvoorbeeld dat De Rijk had geknoeid met bankmachtigingen om zich zo toegang te verschaffen tot het geld van het echtpaar. De Rijk loog aantoonbaar over zulke aangelegenheden. Later zou blijken dat hij het echtpaar om het leven had gebracht en op het terrein van de kinderboerderij had begraven.[8] Een perfecte geluidsopname van het gesprek dat Peter R. de Vries met de man had, werd ons door zijn redactie ter beschikking gesteld. Die geluidsopname analyseerden we met onze *high tech* stemanalyseapparatuur. Het apparaat kwam met een wirwar aan meldingen, zoals *excited* en *subject is not sure*, maar gaf bij geen van de aantoonbare leugens de kwalificatie *probable lie* of *deception*. Dus ook hier een *truth bias*, dit keer in de richting van een zware crimineel die evident zat te liegen.

De resultaten van ons onderzoek staan haaks op de pretenties van de fabrikanten die stemanalyseapparatuur op de markt brengen. Dat voor hen zulke ontnuchterende resultaten niet helemaal onverwacht zullen komen blijkt als men de *License Agreement* van de apparatuur bestudeert. In de *License Agreement* van ons apparaat melden de kleine letters: 'The customer further acknowledges, that the Product is not an automatic lie detector, that no conclusion should be drawn on the basis of the Products' results alone, and that such results should only be referred to in conjunction with other methods, data, and information.' Dat lijkt toch meer op een zonnebank waarvan de handleiding zegt dat je pas bruin wordt als je in de zon gaat liggen. Niet kopen dus.

Noten

1. Holman, D. (2005). Nothing but the truth. *The American Spectator*, 15-12-2005.
2. Sommige fysiologen betwijfelen of de micro-tremoren die in de spieren van ledematen worden geregistreerd ook zijn aan te treffen in de laryngale spiergroepen. Zie Shipp & Izdebski (1981).
3. Grote leveranciers zijn het Amerikaanse *National Institute for Truth Verification*, dat de *Computer Voice Stress Analyzer* op de markt brengt en het Israëlische *Nemesysco* dat onder andere de *Layered Voice Analysis* (LVA) en de *Gate Keeper* (GK1) produceert.
4. Voor wie dit te veel geld en moeite vindt: u kunt uw geluidsfragmenten ook *on line* laten analyseren. Zie bijvoorbeeld http://www.personae.nl/analyse.php
5. Bij dit experiment werden wij geholpen door de volgende studenten: Ine Beaulen, Michael van Damme, Eefke van den Heuvel, Kimberly Kipigroch, Helene Moonen, Maaike Roubroeks en Esther Zeijen. Wij zijn hen daarvoor zeer dankbaar.
6. Onze proefpersonen beoordeelden de filmfragmenten als 'waar' of 'onwaar' en

gaven daarbij ook een zekerheidsschatting (0-100%).
Door alle zekerheidsschattingen voor als 'onwaar' geclassificeerde fragmenten met -1 te vermenigvuldigen, verkrijgt men beoordelingen die lopen van -100 (zeker onwaar) tot 100 (zeker waar). Indien ieder oordeel door 2 wordt gedeeld en er dan 50 bij op wordt geteld, gaat (ook) deze schaal van 0 (zeker onwaar) tot 100 (zeker waar) lopen. Dat maakt het mogelijk om het oordeel van de proefpersonen direct te vergelijken met dat van het stemanalyse-apparaat.

7. Zo tekende *De Telegraaf* uit de mond van een employee van de *Bank of Scotland* op: 'Onze klanten krijgen vooraf te horen dat tijdens het telefoongesprek een leugendetector meeluistert. Binnen vijftien minuten kan het systeem aangeven of zij de waarheid hebben verteld.' *De Telegraaf* vervolgt: 'De grootste verzekeraar van ons land, Centraal Beheer, laat in een reactie weten het nieuwe systeem te willen bestuderen.' *De Telegraaf*, 16-8-2003. In dat verband meldde *de Volkskrant* dat een aantal Engelse verzekeraars gebruikmaakt van de apparatuur om fraudeurs te ontmaskeren. 'De maatschappijen maken gewag van een daling van 20 procent in het aantal valse claims.' *de Volkskrant*, 1-2-2005.
8. Zie voor het volledige dossier: http://www.peterrdevries.nl/

Literatuur

DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J.J. & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, 1, 346-357.

DePaulo, B.M. (2003). Cues to deception. *Psychological Bulletin*, 129, 74-118.

Gamer, M., Rill, H.G., Vossel, G. & Gödert, H.W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology*, 60, 76-87.

Krapohl, D.J., Ryan, A.H. & Shull, K.W. (2002). Voice stress devices and the detection of lies. *Polygraph*, 31, 43-48.

Lykken, D.T. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. New York: Plenum Press.

Merckelbach, H. en Boelhouwer, J. (2002). Leugendetectie. In P.J. van Kopppen, D.J. Hessing, H.L.G.J. Merckelbach en H.F.M. Crombag (Red.). *Het recht van binnen: Psychologie van het recht* (p. 649-666). Deventer: Kluwer.

Merckelbach, H. en Jelicic, M. (2005). *Hoe een CIA agent zijn geheugen hervond en andere waargebeurde verhalen*. Amsterdam; Contact.

Shipp, T. & Izdebski, K. (1981). Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Science*, *26*, 501-505.