Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie

Vakbijlage

Automatische sprekervergelijking

Inhoudsopgave

Inleiding

In vergelijkend spraakonderzoek wordt meestal gevraagd een opname van een onbekende stem te vergelijken met een opname van een bekende stem om uit te zoeken of de twee stemmen van dezelfde persoon zijn of van twee verschillende personen. De opname van de onbekende stem (vaak van de dader) noemen we het betwiste materiaal en de opname van de bekende stem (vaak van de verdachte) noemen we het vergelijkingsmateriaal. Hieronder wordt de methode toegelicht waarbij software gebruikt wordt om te vergelijken. De andere methode, waarbij de mens luistert en vergelijkt, wordt toegelicht in de vakbijlage 'Vergelijkend spraakonderzoek'.

De onderzoeksvraag die gesteld wordt komt meestal neer op: "Is de spreker van het betwiste materiaal dezelfde spreker als die van het vergelijkingsmateriaal?". Bij het NFI vertalen we die vraag naar twee hypothesen:

Zelfde-sprekerhypothese:
Het betwiste materiaal en het vergelijkingsmateriaal zijn afkomstig van dezelfde spreker.

Verschillende-sprekerhypothese:
Het betwiste materiaal en het vergelijkingsmateriaal zijn afkomstig van verschillende sprekers.

Het resultaat van vergelijkend spraakonderzoek drukt uit hoeveel beter de bevindingen passen bij de ene hypothese dan bij de andere.

Vergelijking met automatische sprekervergelijkings-software houdt de volgende stappen in:

1. De vergelijking van de twee audiobestanden door de software
De software vergelijkt het betwist materiaal met het vergelijkingsmateriaal en drukt de mate van overeenkomst uit in een zaakscore.

2. De vergelijkingen van testsprekers
Opnamen van testsprekers uit een NFI-verzameling worden vergeleken door de software. Dat levert zelfde-sprekerscores en verschillende-sprekerscores op.

3. Controle prestatie software
Met de zelfde-sprekerscores en verschillende-sprekerscores wordt gecontroleerd of de software goed genoeg werkt in de omstandigheden van de zaak.

4. Interpretatie zaakscore en bepaling bewijskracht
De zaakscore wordt afgezet tegen de zelfde-sprekerscores en verschillende-sprekerscores om de bewijskracht van de zaakscore te bepalen.

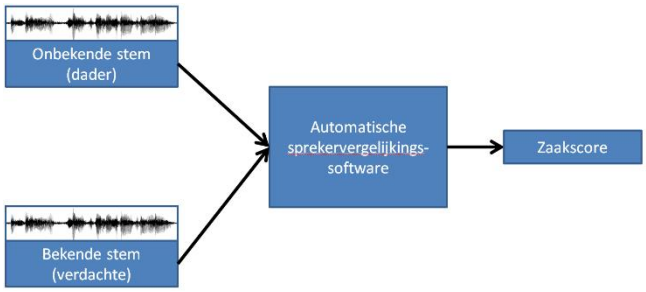
5. Controle robuustheid
Het hele onderzoek wordt een aantal keer opnieuw uitgevoerd met telkens net andere keuzes in de onderzoeksopzet. Daarmee wordt gecontroleerd of de uitkomst van het onderzoek robuust is voor kleine veranderingen in de gemaakte keuzes.

6. Naar de eindconclusie
Het resultaat van de automatische sprekervergelijking wordt samengenomen met de uitkomst van de methode waarbij de mens luistert en vergelijkt.

1. De vergelijking van de twee audiobestanden door software

Gebruikte software
De software die op het NFI gebruikt wordt om te vergelijken is momenteel (2021) Vocalise 2019, 2.7.0.1650 van Oxford Wave Research. Van deze software is uit validatie-onderzoek gebleken dat die in principe geschikt is voor gebruik in NFI-zaakonderzoek. Als uit validatie-onderzoek in de toekomst blijkt dat andere software beter presteert, kan het gebeuren dat het NFI andere software zal gaan gebruiken. Dat zal naar verwachting geen grote gevolgen hebben voor hoe de software in zaakonderzoek wordt ingezet.

Werking van de software
De opname met het betwiste materiaal en de opname met het vergelijkingsmateriaal worden aangeboden aan de software. De software haalt uit elk van de opnamen kenmerken uit de spraak. Deze kenmerken beschrijven grofweg de klank van de stem, en ze zijn dus niet afhankelijk van wat er precies gezegd wordt in de opnamen. Deze kenmerken worden – binnen de software – numeriek vastgesteld.
Vervolgens gebruikt de software de kenmerken van het betwiste materiaal en de kenmerken van het vergelijkingsmateriaal om een mate van overeenkomst te bepalen en meet die in een score. Deze score drukt uit in hoeverre de stemmen uit de twee opnamen op elkaar lijken. Deze score – dus tussen het betwiste materiaal en het vergelijkingsmateriaal – noemen we hier de zaakscore, omdat het de score is die uit de vergelijking van het zaakmateriaal komt.



Figuur 1. Werking van automatische sprekervergelijkingssoftware. Twee opnamen van spraak worden ingeladen in de software. Die bepaalt een mate van overeenkomst tussen de opnamen van stemmen, en drukt die uit in de zaakscore.

De zaakscore
De zaakscore is een getal dat meestal tussen de -5 en 5 ligt. Het getal zelf heeft geen betekenis anders dan: hoe hoger, des te meer overeenkomst er is gevonden door de software. Om dat getal goed te kunnen interpreteren als bewijs dat richting de zelfde-sprekerhypothese wijst of richting de verschillende-sprekerhypothese, moet dit getal nog worden afgezet tegen scores waarvan bekend is of die afkomstig zijn uit vergelijkingen van dezelfde spreker of uit vergelijkingen van verschillende sprekers.

Meer zaakmateriaal
In veel zaken is er meer dan één opname betwist materiaal of meer dan één opname vergelijkingsmateriaal. Bij meer dan één opname vergelijkingsmateriaal worden die zoveel mogelijk gezamenlijk aan de software aangeboden, zodat die een zo compleet mogelijk beeld van de stem van de spreker kan gebruiken. Als er meer dan genoeg vergelijkingsmateriaal is, kan het voorkomen dat de onderzoeker een aantal opnamen uit het vergelijkingsmateriaal kiest. Daarbij wordt gekeken naar de hoeveelheid spraak in de opname en hoe goed de omstandigheden waaronder de spraak is opgenomen overeenkomen met de omstandigheden in het betwiste materiaal.

Omdat van het betwiste materiaal niet bekend is wie er spreekt, kan in zaken met meerdere betwiste opnamen meestal niet worden aangenomen dat het allemaal van dezelfde spreker is. Daarom worden de betwiste opnamen apart behandeld, dus in een zaak met vijf betwiste opnamen wordt het gehele onderzoek vijf keer uitgevoerd: eenmaal voor elke betwiste opname.

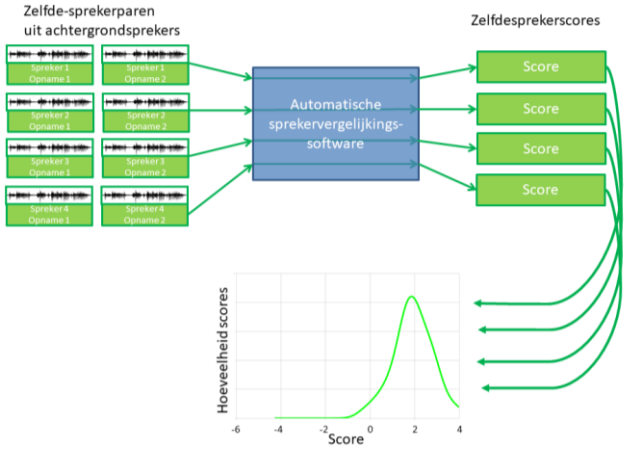
2. De vergelijkingen van testsprekers

Opname-omstandigheden
De zaakscore uit een vergelijking met software van het zaakmateriaal is natuurlijk afhankelijk van hoeveel de stemmen in de opnamen op elkaar lijken. Maar: eigen validatie-onderzoek en de wetenschappelijke literatuur op dit punt laten zien dat de uitkomst ook wordt beïnvloed door andere factoren. Voorbeelden daarvan zijn het opnametype (telefoontap / verhoor / opname vertrouwelijke communicatie), de hoeveelheid spraak in een opname en de gesproken taal in de opname.
Omdat er meer dan alleen sprekerkenmerken van invloed zijn op de zaakscore, kan de zaakscore niet op zichzelf geïnterpreteerd worden. Wat in een vergelijking tussen een telefoontap en een verhoor een hoge score is, kan in een vergelijking tussen twee telefoontaps een lage score zijn. Alleen de zaakscore is dus niet genoeg: die moet worden bekeken in het licht van hoe de software normaalgesproken scoort bij vergelijkingen in de omstandigheden van de zaak.

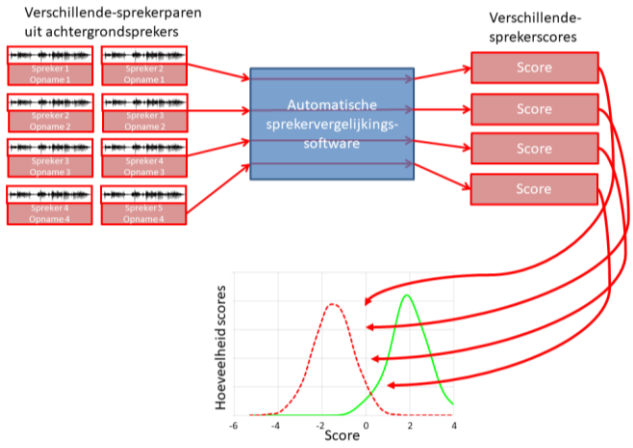
Selectie testsprekers
Om vergelijkingen in de omstandigheden van de zaak mogelijk te maken wordt gebruik gemaakt van opnamen van testsprekers. Die opnamen komen uit een verzameling forensisch realistisch audiomateriaal die door het NFI is aangelegd. In die verzameling is bekend wie welke spreker is. Het is dus mogelijk om daar paren van opnamen uit te kiezen waarvan bekend is dat daarin dezelfde spreker voorkomt (zelfde-sprekerpaar) en paren van opnamen waarvan bekend is dat het om twee verschillende sprekers gaat (verschillende-sprekerpaar).
Deze paren van opnamen worden geselecteerd op de omstandigheden van de zaak. In een zaak kan het bijvoorbeeld gaan om het vergelijken van twee telefoontaps, gesproken in het Nederlands, waarbij de betwiste opname bestaat uit 30 seconden spraak en het vergelijkingsmateriaal uit 45 seconden spraak. Paren van opnamen met diezelfde kenmerken worden uit de verzameling geselecteerd en zo nodig geknipt om de juiste opnameduur te bereiken.

Vergelijking testsprekers
Er zijn dus twee soorten paren van opnamen: zelfde-sprekerparen en verschillende-sprekerparen. Deze paren opnamen worden allemaal vergeleken door de software, op dezelfde manier zoals dat al gebeurde met het zaakmateriaal. Dat levert twee soorten scores op: scores uit vergelijkingen van een zelfde-sprekerpaar (zelfde-sprekerscores) en scores uit vergelijkingen van een verschillende-sprekerpaar (verschillende-sprekerscores).

Zelfde-sprekerscores en verschillende-sprekerscores
Deze twee typen scores worden gevisualiseerd in twee scoredistributies. Zie figuren 2 en 3.



Figuur 2. Paren van opnamen waarvan zeker is dat die dezelfde spreker bevatten (zelfde-sprekerparen) worden onderling vergeleken. Daaruit komen zelfde-sprekerscores, die worden weergegeven in een distributie (de groene lijn in de grafiek). In dit voorbeeld liggen de meeste zelfde-sprekerscores rond 2.



Figuur 3. Paren van opnamen waarvan zeker is dat die verschillende sprekers bevatten (verschillende-sprekerparen) worden onderling vergeleken. Daaruit komen verschillende-sprekerscores, die worden samengevat in een distributie (de rode lijn in de grafiek). In dit voorbeeld liggen de meeste verschillende-sprekerscores rond -2.

De zelfde-sprekerscoredistributie (de groene curve) bestaat uit scores die waargenomen zijn als de zelfde-sprekerhypothese waar is. De verschillende-sprekerdistributie (de rode curve) bestaat uit waargenomen scores als de verschillende-sprekerhypothese waar is. Voor de vergelijking met het zaakmateriaal vormen de groene en rode curve de scores die we kunnen verwachten als het een zelfde-sprekerpaar is of een verschillende-sprekerpaar.
Met deze twee distributies worden twee dingen gedaan:
Ten eerste wordt gekeken of de software wel goed presteert in de omstandigheden van de zaak (paragraaf 3 hieronder).
Ten tweede kan worden bekeken of de zaakscore beter past bij de zelfde-sprekerscores of bij de verschillende-sprekerscores (paragraaf 4 hieronder).

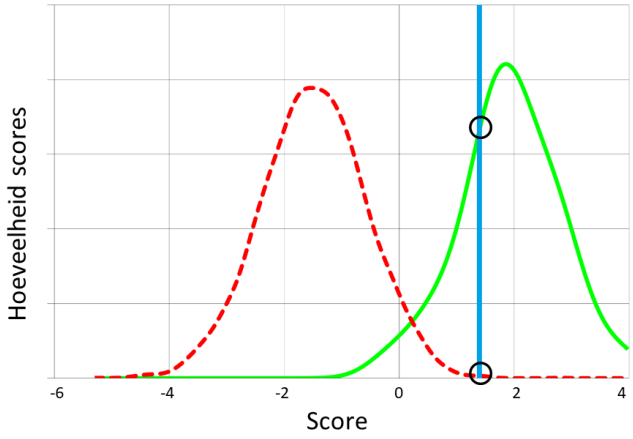
3. Controle prestatie software

Om te controleren of de software voldoende onderscheid kan maken tussen sprekers in de omstandigheden van de zaak, wordt gekeken naar de twee scoredistributies.
Als de twee distributies ver uit elkaar liggen, betekent dat dat de software in de omstandigheden van de zaak goed onderscheid kan maken tussen zelfde-sprekerparen en verschillende-sprekerparen. Zelfde-sprekervergelijkingen leveren dan immers meestal hogere scores dan verschillende-sprekerparen, en de uitkomst van de software bij de zaakvergelijking geeft dan waarschijnlijk bruikbare informatie.
Als de twee distributies voor een groot deel overlappen, betekent dat dat de software in de omstandigheden van de zaak niet goed onderscheid kan maken. Of er nu tweemaal dezelfde spreker of twee verschillende sprekers worden vergeleken: het levert altijd ongeveer dezelfde score op. De zaakscore levert dan dus geen bruikbare informatie op. Als dit het geval blijkt in een zaak, zal in het rapport kortweg worden vermeld dat uit onderzoek is gebleken dat het materiaal zich niet leent voor automatische sprekerherkenning en zal de methode verder niet worden toegepast.

4. Interpretatie zaakscore

Met deze twee distributies kan bovendien de zaakscore geïnterpreteerd worden. Er kan nu namelijk gekeken worden of de zaakscore meer voorkomt bij zelfde-sprekervergelijkingen of juist meer voorkomt bij verschillende-sprekervergelijkingen. In Figuur 4 is hiervan een voorbeeld gegeven. De zaakscore van ongeveer 1.5 (aangegeven met de blauwe lijn) kruist de groene distributie op een veel hoger punt dan de rode distributie. Dat betekent dat een score van 1.5 veel meer voorkomt bij zelfde-sprekervergelijkingen dan verschillende-sprekervergelijkingen. Dit is een resultaat dat veel beter past bij een zelfde-sprekervergelijking dan bij een verschillendesprekervergelijking. Dit zou zich dus vertalen in een bewijskracht die naar de zelfde-sprekerhypothese wijst. Als de zaakscore op -2 zou liggen, is dat juist andersom. Die zaakscore komt veel meer voor bij verschillende-sprekervergelijkingen dan bij zelfde-sprekervergelijkingen en het resultaat zou juist meer naar de verschillende-sprekerhypothese wijzen.
Hoeveel beter de zaakscore past bij een zelfde-sprekervergelijking dan bij een verschillende-

sprekervergelijking kan numeriek worden uitgedrukt in een likelihood ratio, zie de toelichting bij figuur 4.



Figuur 4. De scores uit zelfde-sprekervergelijkingen (groene lijn) en de scores uit verschillende-sprekervergelijkingen (rode, onderbroken lijn) worden gebruikt om de bewijskracht van de zaakscore (verticale blauwe lijn) te bepalen. In dit voorbeeld past de zaakscore veel beter bij een zelfde-sprekerscore dan bij een verschillende-sprekerscore. De zaakscore (van ca. 1.5) kruist de groene lijn in dit voorbeeld op een hoger punt dan de rode lijn, wat betekent dat de zaakscore veel meer voorkomt bij zelfde-sprekervergelijkingen dan bij verschillende-sprekervergelijkingen. De likelihood ratio (de maat voor bewijskracht) wordt berekend door de hoogtes van het kruispunt van de blauwe en groene lijn en van het kruispunt van de blauwe en rode lijn (aangegeven met cirkels) op elkaar te delen.

5. Controle robuustheid

In het onderzoek zijn een aantal stappen gezet die op een inschatting van de onderzoeker berusten. Om te testen of de methode robuust is voor dit soort inschattingen, wordt het hele onderzoek herhaald, waarbij op die punten net andere keuzes worden gemaakt. Het aantal testsprekers wordt bijvoorbeeld wat verkleind, of er wordt een andere selectie van vergelijkingsmateriaal gebruikt.
Als de uitkomsten daarvan in de buurt liggen van de oorspronkelijke uitkomst, is de methode kennelijk robuust voor deze inschattingen. In dat geval wordt de ordegrootte van al die uitkomsten gebruikt in het bepalen van de eindconclusie.
Als de uitkomsten daarvan sterk onderling verschillen, dan blijkt dat de methode erg gevoelig is voor de keuzes die de onderzoeker maakt in de opzet. In dat geval wordt de uitkomst van het onderzoek met meer voorzichtigheid behandeld.

6. Naar de eindconclusie

In de meeste zaken loopt de automatische sprekervergelijking parallel met een vergelijking uitgevoerd door de mens. De uitkomsten van deze twee methoden worden samengenomen tot één eindconclusie. Afhankelijk van de omstandigheden in de zaak kan de ene methode zwaarder worden meegewogen dan de andere. Per zaak wordt bekeken hoe deze uitkomsten zich vertalen naar één eindconclusie.
Daarbij wordt rekening gehouden met:

- hoe geschikt het zaakmateriaal is voor elk van de methodes
- hoe goed de opnameomstandigheden van de testsprekers overeenkomen met de opnameomstandigheden in het zaakmateriaal
- hoe goed de spraakonderzoeker in de menselijke methode bekend is met het type spraakmateriaal in de zaak
- hoe robuust de methode bleek uit de verschillende uitvoeringen (zie paragraaf 5)
- hoe goed elk van de methoden getest en gevalideerd is.

De eindconclusie wordt uitgedrukt in een verbale schaal, waarin wordt aangegeven hoeveel waarschijnlijker de bevindingen zijn onder de ene hypothese dan onder de andere. Zie voor meer uitleg over de conclusieschaal de vakbijlage 'Vergelijkend spraakonderzoek'.

Voor algemene vragen kunt u contact opnemen met de Frontdesk, telefoon (070) 888 68 88. Voor inhoudelijke vragen kunt u contact opnemen met het onderzoeksgebied Spraak- en Audio-onderzoek van de afdeling Digitale en Biometrische Sporen, telefoon (070) 888 6425.

Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie
Postbus 24044 | 2490 AA Den Haag

Telefoon (070) 888 66 66
www.forensischinstituut.nl

18 december 2019