Een kopieermachine voor je stem

Zijn computerstemmen nog wel van echte te onderscheiden?

Deze publicatie is onderdeel van het thema Over taal gesproken op Kennislink.nl.

Computers praten al tientallen jaren, al gaat dat vaak met een wat metaalachtige klank. De laatste tijd verandert dit: stemmen worden steeds sneller geleerd én geïmiteerd. Kunnen neurale netwerken van de computer een ware meesterverteller maken?

Door Roel van der Heijden en Erica Renckens

De Amerikaanse president Donald Trump en zijn voorganger Barack Obama hebben het gezellig samen. Ze keuvelen over een nieuwe technologie waarmee een computer stemmen razendsnel kopieert. De Amerikaanse politica Hillary Clinton maakt het onderonsje compleet als ze bijspringt om te vertellen hoe het ongeveer werkt.

Het korte fragment is geproduceerd door het Canadese bedrijf Lyrebird, dat beweert dat hun stemcomputer slechts enkele korte fragmenten van de Amerikaanse politici nodig had om hun stemmen te kopiëren. Ze maakten daarbij gebruik van zogeheten neurale netwerken, een techniek die in zekere zin lijkt op de manier waarop onze hersenen informatie verwerken en opslaan.

Hoewel het meteen duidelijk is dat we niet naar de echte Trump, Obama en Clinton luisteren, is de snelheid waarmee dit algoritme leert ongeëvenaard. Een indrukwekkende ontwikkeling op het gebied van het namaken van stemmen lijkt aanstaande. Dat heeft handige toepassingen, voor bijvoorbeeld mensen die hun stem dreigen te verliezen. Maar een gekopieerde stem geeft kwaadwillenden ook kansen. Moeten we ons zorgen maken en hoe gaat dit klonen van een stem precies in zijn werk?

Computers leren praten

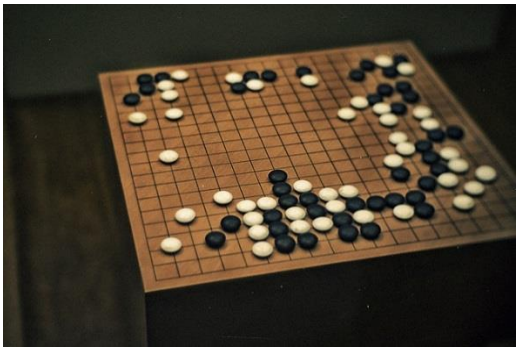
Pratende computers, zoals het navigatiesysteem in je auto of de omroepberichten op het station, kennen we al langer. Deze systemen gebruiken spraakopnames van een stemacteur, waaruit hele woorden of alleen bepaalde woorddelen geselecteerd worden. Door de verschillende fragmenten aan elkaar te plakken, kun je de computer alles laten zeggen wat je maar wilt, al blijf je horen dat het geen natuurlijke spraak is.

Die opnames creëren is monnikenwerk. Om vrijwel iedere mogelijke klank vast te leggen, moet de stemacteur een enorm aantal speciaal geselecteerde zinnen inspreken. Esther Klabbers is onderzoeker bij ReadSpeaker, een bedrijf dat verschillende digitale stemmen in acht talen aanbiedt voor bijvoorbeeld een 'voorleesknop' op websites. Hun stemacteurs spreken soms wel twintigduizend zinnen in. "Uiteindelijk creëer je een database met honderdduizend tot tweehonderdduizend zogenoemde 'spraakonderdelen' die kunnen bestaan uit woorden, lettergrepen of een enkele klank", zegt Klabbers. "Daarnaast is er een uitgebreid 'woordenboek' nodig waarin staat hoe woorden worden uitgesproken. Op basis daarvan kiest de computer spraakonderdelen en plakt ze zo goed mogelijk aan elkaar."

Veel elektronische stemmen klinken nogal eentonig en emotieloos. Dat komt omdat de opnames waarop ze gebaseerd zo monotoon mogelijk zijn ingesproken. De computer kan audiofragmenten dan makkelijker aan elkaar plakken. *lincolnblues via CC BY-NC-ND 2.0*

Net als in het brein

Maar het kan ook anders, zoals de demonstratie van Lyrebird laat zien. Hoewel details over het algoritme van de Canadese start-up ontbreken, is duidelijk dat ze gebruik maken van neurale netwerken. Een fundamenteel andere aanpak, die een beetje lijkt op hoe ons brein leert. Het gaat uit van een algoritme dat data analyseert, verwerkt en daar vervolgens van leert. Door te kijken naar de eigen fouten – als het de output vergelijkt met de originele stem – perfectioneert het programma een stem, zonder dat daar nog een programmeur voor nodig is.



Neurale netwerken worden met succes in verschillende vakgebieden ingezet. Computers leren zo inmiddels autorijden (Tesla), ze verslaan professionele Go-spelers (AlphaGo) of zijn in staat objecten op foto's te herkennen.
Linh Nguyen via CC BY-NC-ND 2.0

Er is dus nog steeds een trainingsopname nodig. De computer 'luistert' ernaar en bepaalt voor ieder moment in de opname een paar honderd parameters, zoals de uitgesproken klanken, de plek van de klemtoon en de timing van de spreker. Het algoritme bepaalt zo welke parameters er bij specifieke woorden en zinsdelen in een tekst horen. Bij de reproductie van een stem kiest de computer op basis van wat hij leerde nieuwe parameters en maakt daar een gesproken stem van.

Het voordeel van deze aanpak is dat een computer een stem veel sneller kan leren dan met de 'ouderwetse' spraaksynthese. Klabbers schat dat een goed neuraal netwerk genoeg heeft aan een tiende van de opnamesessies via de oude methode. Ook zegt ze dat de gecreëerde stem veel flexibeler is. "Als je bijvoorbeeld een meer expressieve spraak wil, dan pas je simpelweg de spraakparameters aan", aldus Klabbers. "Dat is bij de oude methode niet mogelijk."

Een nadeel is de relatief grote rekenkracht die een neuraal netwerk vergt. Klabbers vertelt over de spraakgenerator WaveNet van Google, die volgens een vergelijkbaar principe met een groot aantal spraakparameters werkt. "Dat gaat goed, maar voor het trainen en genereren van de stemmen is een flinke computer nodig", zegt ze. "Een zin reconstrueren kan op dit moment wel negentig seconden duren.

Het is daarom een methode die nog niet geschikt is om bijvoorbeeld op je mobiele telefoon te gebruiken." Dit probleem verdwijnt natuurlijk met snellere computers en betere technieken. De resultaten laten in ieder geval zien dat de techniek veelbelovend is. Ook andere bedrijven, waaronder ReadSpeaker, doen onderzoek naar het genereren van stemmen met neurale netwerken.

Unieke stemmen

Elke stem is uniek. Misschien niet net zo uniek als je DNA of je vingerafdruk – goede stemimitators als Jochem Myjer kunnen bekende stemmen overtuigend nadoen. Maar wel zo bijzonder dat je je vrienden met je ogen dicht kunt herkennen, puur op hun spraakgeluid.

Die persoonlijke stem wordt onder andere gevormd door je stembanden, die in je strottenhoofd de lucht uit je longen in beweging kunnen brengen. De lengte en de dikte van je stembanden bepalen hoe snel ze trillen en daarmee hoe hoog (of laag) je stemgeluid is. Dikke en lange stembanden laten de lucht uit je longen langzaam trillen en zorgen zo voor een lage klank. De snelle luchttrillingen uit dunne en korte stembanden geven juist een hoog geluid.

De in trilling gebrachte lucht resoneert vervolgens in je borst- en hoofdholtes. Die zijn bij iedereen net iets anders gevormd en zorgen zo voor een uniek stemgeluid. Met de spieren in en rondom je mond kun je je mondholte vervormen en zo verschillende klanken creëren. Ook dit doet iedereen op zijn eigen, unieke manier, wat samen met veranderingen in toonhoogte en spreeksnelheid zorgt voor een onderscheidende spreekstijl.

Je brein voor de gek

Het fragment dat Lyrebird recentelijk de wereld in slingerde, klinkt indrukwekkend, maar is nog duidelijk te herkennen als nep. De intonatie van de sprekers klinkt niet altijd natuurlijk en ook hebben de stemmen nog iets blikkerigs. De start-up zegt zelf deze onvolkomenheden op korte termijn op te lossen, waarna de kunstmatige spraak niet meer van echt te onderscheiden zal zijn.

Bewerkte foto's kennen we al. Ze kunnen soms overtuigend zijn en grote gevolgen hebben, zoals deze door Wilders gefotoshopte foto van Pechtold tussen Hamas-aanhangers. Maar wat nu als we ook spraakopnames of zelfs video's overtuigend zouden kunnen manipuleren? Kun je iemand kunstmatig woorden in de mond leggen? *Geert Wilders*

Spraakonderzoeker James McQueen gelooft niet dat het in zo'n vaart zal lopen. "Zeker niet op basis van één minuut opgenomen spraak", reageert de hoogleraar Spraak en Leren aan de Radboud Universiteit. "Als je spreekt gebruik je zoveel verschillende klanken. Om daar een compleet repertoire van te krijgen, heb je veel meer data nodig. Neem bijvoorbeeld mijn r-klank. Meestal gebruik ik de typisch Zuid-Engelse variant achter in de keel, maar soms schiet er nog een rollende 'r' tussendoor, die mijn Schotse achtergrond verraadt. Dat is karakteristiek voor mijn spraak en dat moet je dan maar net gevangen hebben in die ene minuut."

Sowieso is ons brein een pietje-precies als het aankomt op luisteren naar spraak. "Doordat iedereen net iets anders spreekt, moeten je hersenen zich steeds aanpassen aan een nieuwe spreker. Dat *intunen* gaat vanzelf; ons brein is heel adaptief. Uit onderzoek blijkt dat we zelfs een geheugen hebben voor de uitspraak van klanken door specifieke sprekers. Kleine afwijkingen vallen direct op", zegt McQueen.

Ook Klabbers is niet zo bang dat spraakcomputers binnenkort niet meer te onderscheiden zijn van echte stemmen. Van de snelheid waarmee Lyrebird stemmen leert is ze onder de indruk. “Maar ik vraag me af of mensen nu voor zo’n product willen betalen, zo goed zijn de stemmen nog niet”, zegt ze. “En ook al is een leek misschien sneller overtuigd, als je een getraind oor hebt, dan haal je kunstmatige stemmen er zo uit.”

Als je de techniek van Lyrebird combineert met videomanipulatie zijn de mogelijkheden helemaal indrukwekkend – en een beetje beangstigend. De Duitse universiteit van Erlangen, het Max Planck Instituut voor Informatica en Stanford University ontwikkelden samen Face2Face, een methode om de gezichtsbewegingen in video te manipuleren. De bewegingen van een acteur kunnen hierbij als een soort masker over een gezicht geplaatst worden. “Als je alleen audio hebt, richt je daar al je aandacht op en merk je foutjes eerder op”, zegt McQueen. “Maar visuele input draagt ook bij aan wat je verstaat, dus zo’n combinatie van beeld en geluid zou eventuele tekortkomingen van de spraaksynthese kunnen maskeren.”

Fotoshop voor stemmen

Maar de techniek maakt snelle sprongen: stel dat ook een spraakexpert het verschil tussen kunstmatig en echt niet meer hoort. Misschien dat de computer dan zélf nog uitkomst biedt. Alle factoren die bijdragen aan een stemsignaal (zoals vibratie van de stembanden, weerkaatsingen in de mond, keel en hoofdholtes, intonatie) zijn nauwkeurig te analyseren. “De computerstem zal waarschijnlijk veel regelmatigere patronen laten zien dan een natuurlijke stem”, aldus Klabbers.

Het wordt moeilijker om echte van kunstmatige stemmen te onderscheiden als de opnamekwaliteit lager is, zoals bij een stem over een telefoonlijn.
Pixabay, Olly Browning via CC0

Leven we uiteindelijk in een wereld waarin opnames helemaal niet meer te controleren zijn? Kun je familie of vrienden nog wel vertrouwen als je ze aan de telefoon hebt? “Het klonen van stemmen heeft wel risico’s”, geeft Klabbers toe. “Maar vergelijk het met een enigszins vergelijkbare techniek: photoshop. Inmiddels weten mensen dat ook foto’s gemanipuleerd kunnen zijn, misschien dat we ook leren leven met hetzelfde idee voor stemmen.”