Spraaktechnologen willen sprekende computers zó natuurlijk laten klinken dat je bijna geen verschil hoort met mensen. Nu zijn we ongeveer op dat punt beland. Missie geslaagd?

Erica Renckens Josje van Koppen

Pratende computers: **beangstigend goed**

"Het voortbestaan van Nederland staat op het spel." Met deze dreigende woorden licht premier Rutte tijdens een officiële toespraak vanuit het Torentje zijn plannen voor de aanpak van de klimaatcrisis toe. "Het Nederlandse bedrijfsleven wordt met genereuze overheidssteun toekomstbestendig gemaakt. Wie toch de hakken in het zand zet, moet de deuren sluiten. Dit geldt ook voor bedrijven in de industrie en de landbouw."

Behoorlijk schokkende en onverwachte woorden uit de mond van de Nederlandse neoliberale premier. Het filmpje van de toespraak is dan ook nep: hij heeft deze woorden nooit uitgesproken. We kijken naar een deepfake van het journalistieke platform *De Correspondent*, waarin met behulp van kunstmatige intelligentie beeld en geluid zijn gecreëerd. De audiovisuele versie van nepnieuws – nauwelijks van echt te onderscheiden.

Een belangrijk aspect van de deepfake is de stem; die moet onmiskenbaar klinken als die van de persoon die in het filmpje te zien is. "Tijdens zo'n monoloog komen we inmiddels heel dicht bij perfectie", vertelt Max Louwerse, hoogleraar cognitieve psychologie en kunstmatige intelligentie aan Tilburg University. "Maar in een dialoog valt een computerstem al snel door de mand. Daar komt veel meer expressie bij kijken. Verder kan een stem wel heel erg lijken op die van de echte spreker, maar als je direct gaat vergelijken met het origineel hoor je toch nog verschillen. Dat is nog een stap verder."

Staccato en monotoon

Synthetische stemmen komen we niet alleen tegen in deepfakes, ze worden al sinds de jaren negentig gebruikt in allerlei dagelijkse toepassingen. Van het navigatiesysteem in de auto dat je de weg wijst naar je bestemming tot de omroepinstallatie op het station die je informeert over treinvertragingen en spoorwijzigingen. Van de software die teksten voorleest tot de spraakcomputer die mensen met communicatieproblemen helpt om zich verstaanbaar te maken.

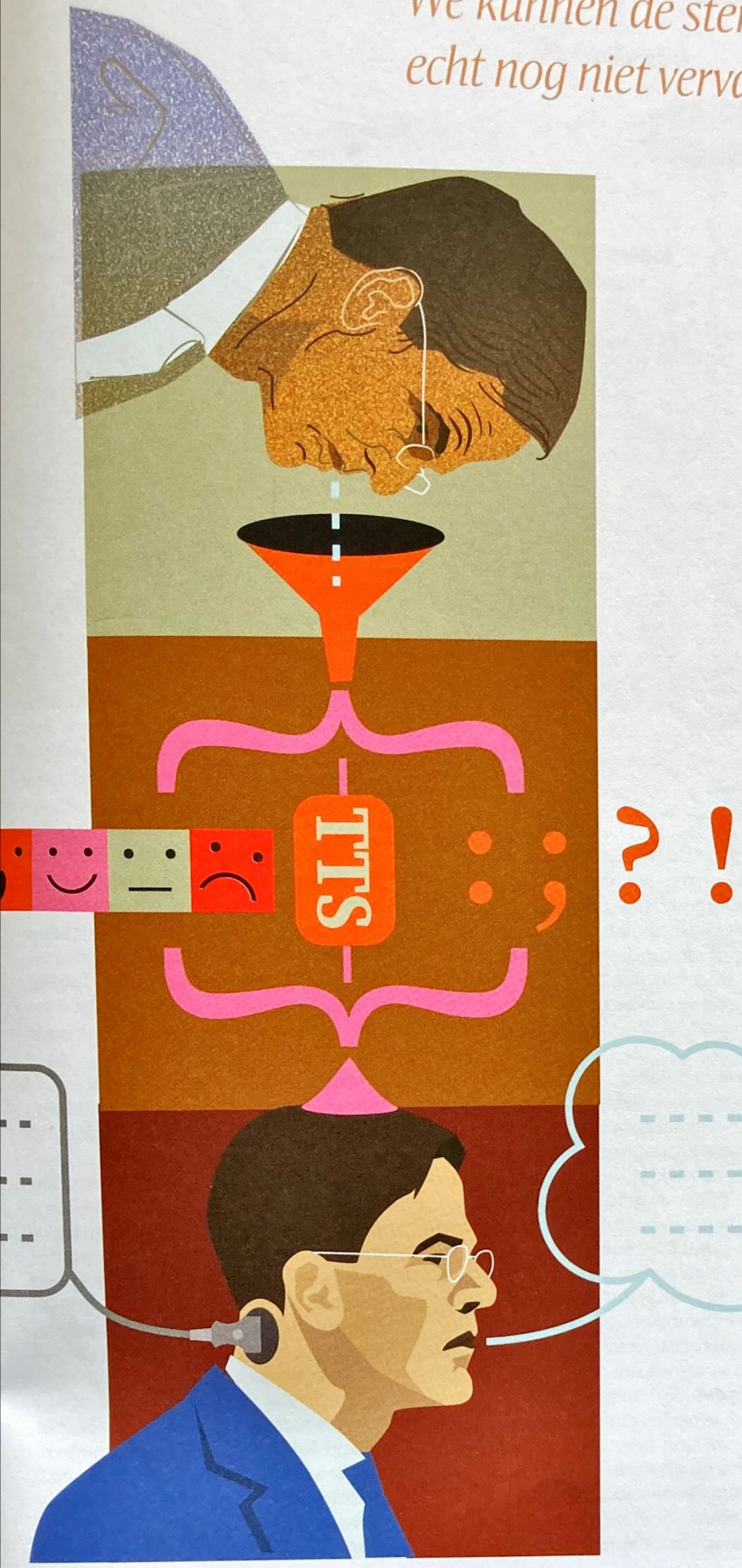
Tot enkele jaren geleden klonken die door de computer gegenereerde stemmen nog staccato en monotoon. De spraak was in feite een reeks aan elkaar geplakte audiofragmenten, wat een hoorbaar onnatuurlijk resultaat gaf. Tegenwoordig gebruiken de zogenoemde text-to-speech-systemen (oftewel TTS-systemen) neurale netwerken. Deze techniek bootst na hoe onze hersenen informatie verwerken: het systeem herkent patronen in grote hoeveelheden data en voorspelt op basis daarvan wat er komt. Als die voorspelling niet blijkt te kloppen, stelt het systeem zijn algoritmes bij, zodat het de volgende keer beter kan inschatten wat er komt. Het resultaat van deze neurale TTS-systemen is soms beangstigend goed.

Het neurale netwerk voor een synthetische stem is getraind met vele uren ingesproken teksten. Op basis daarvan herkent het zelf patronen en legt het relaties tussen woorden en klanken. Zo ontwikkelt het een eigen algoritme dat kan voorspellen hoe een woord wordt uitgesproken – ook als dat niet in het trainingsmateriaal voorkomt. Het algoritme bevat niet alleen het recept voor het stemgeluid, maar ook voor de uitspraak, intonatie en spreeksnelheid.

Binnen vijf seconden

Voor een synthetische stem die alleen is getraind met opnames van één specifieke persoon is zo'n veertig tot tachtig uur aan originele spraak nodig. Maar steeds meer bedrijven werken met een soort moedermodel dat is getraind op een groot aantal stemmen, waarna het algoritme alleen nog maar hoeft te worden gefinetuned met enkele zinnen van de doelstem. Zo werkt Google momenteel aan een neuraal netwerk dat na een training met duizenden stemmen aan zo'n vijf seconden spraak al genoeg heeft om een stem en spreekstijl te imiteren.

“We kunnen de stemacteur voorlopig dus echt nog niet vervangen, gelukkig maar.”



Na de training leest het TTS-systeem vervolgens in drie stappen een tekst voor. Eerst zet een taalkundige module de geschreven tekst om in een representatie van de uitspraak. Zo klinkt in het woord *vader* de *a* als een lange *aa* en de *e* als een zogenoemde stomme *e*. Daarna voorspelt een neuraal akoestisch model welke akoestische kenmerken bij die uitspraak horen. Dit gaat om onder andere timbre, klemtoon, intonatie, adempauzes en spreeksnelheid. Ten slotte gaan deze akoestische kenmerken door een neurale vocoder (een elektronische stemvervormer), die ze vertaalt naar een geluidsgolf die wordt afgespeeld.

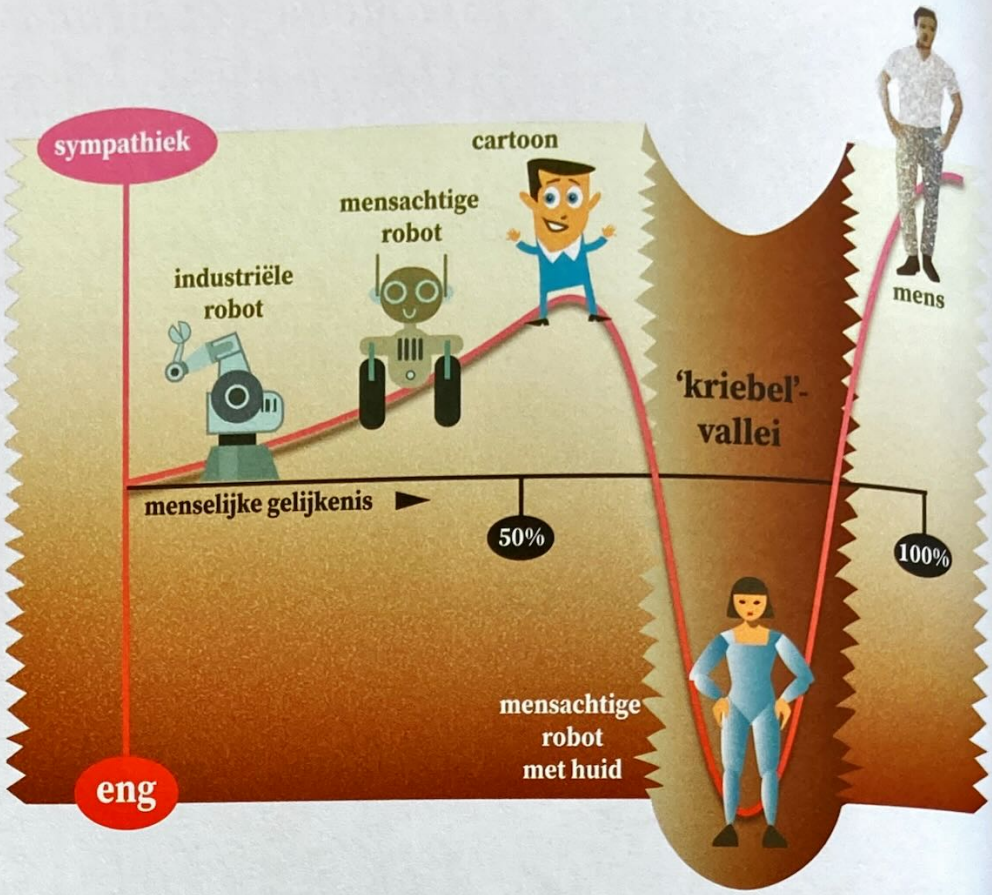
‘Het spijt me’

De computerstemmen die het bedrijf ReadSpeaker ontwikkelt voor met name voorleesfuncties, klinken dankzij deze methode behoorlijk natuurlijk, vertelt spraaktechnoloog Esther Klabbers. “Soms klinkt het nog wel wat saai en monotoon, dus we onderzoeken hoe we meer expressie in de stem kunnen brengen. Dat ligt zowel aan hoe goed het neurale netwerk informatie in zijn model kan verwerken, als aan het trainingsmateriaal. Als je vrolijke, enthousiaste spraak wilt genereren, moet die ook in je originele opnames zitten.”

Nu hoeven nieuwsberichten en zakelijke teksten doorgaans niet opgetogen te worden voorgelezen, maar steeds meer bedrijven gebruiken in hun klantcontact gesproken dialoogsystemen, waarbij de computer in gesprek gaat met de klant. Denk aan een robot-receptionist die de bezoekers de weg wijst naar hun afspraken of een geautomatiseerde telefonische klantenservice die vragen beantwoordt of bellers doorverbindt. Dan speelt expressie een belangrijke rol en dat is momenteel nog een struikelblok voor TTS-systemen.

Klabbers: “Een gesprek begin je eerder met een vriendelijke dan met een neutrale toon. En als een klant →

> *Bedrijven zetten vaak een onzichtbare code in hun synthetische stemmen, zodat je altijd kunt aantonen dat deze nep zijn.*



komt met een klacht, wil je dat 'Het spijt me' uit de computer ook verontschuldigend klinkt. Dat soort zinnen moet je dan dus ook opnemen in je trainingsdata. Vervolgens moet je die verschillende sentimenten ook labelen, zodat het neurale netwerk het verschil ertussen kan maken."

Emotie in de stem

Ook Cerence, een bedrijf gespecialiseerd in virtuele assistenten voor in auto's, werkt aan een natuurlijke spontane uitspraak. Deze assistenten wijzen niet alleen de weg, maar lezen bijvoorbeeld ook nieuws of ontvangen mails voor en maken notities. "We zorgen voor veel variatie in de trainingsdata: vragen, uitroepen, korte zinnen, lange zinnen", vertelt onderzoeksteamleider Jürgen Van de Walle. "Veel van die variatie zie je terug in de interpunctie, met vraagtekens, komma's en dubbele punten. Zo leert het systeem vanzelf al de juiste intonatie."

Hoe weet het systeem dan vervolgens met welke emotie een zin uitgesproken moet worden? Van de Walle: "De toon zit al in de trainingsdata en dus ook in het neurale model. Voor sommige stemmen nemen we spraak bovendien in verschillende stijlen of emoties op. Een ander soort netwerk kan dan bepalen welke stijl of emotie gepast is om te gebruiken in de synthetische spraak."

'Kriebel'-vallei

Hoewel voor beide bedrijven het doel is hun synthetische stemmen zo echt mogelijk te laten klinken, heeft Louwerse daar toch zijn bedenkingen bij. "Vlak vóór de perfecte natuurlijkheid vinden mensen het juist eng; dat is het zogenoemde uncanny-valley-effect", legt hij uit. "Vergelijk het met een robot: hoe menselijker die is, hoe meer je gaat letten op de details en hoe sneller je er de kriebels van krijgt als die niet helemaal kloppen. Tot de robot niet meer van echt te onderscheiden is; dan accepteren we hem weer. Een student van mij heeft dit effect onderzocht voor computerstemmen. Vlak voordat de computerstem helemaal natuurlijk was, vonden mensen het moeilijker om aan te geven of de stem van een mens was of niet. Die langere reactietijd wijst er voorzichtig op dat het uncanny-valley-effect ook geldt voor synthetische stemmen. Als bedrijf zou ik daarom vóór die vallei blijven, als wetenschapper wil ik 'm juist oversteken en die perfectie proberen te bereiken."

Als je een synthetische stem een-op-een gaat vergelijken met het origineel, hoor je nog altijd kleine verschillen. En zeker in dialoogsystemen blijft de natuurlijkheid van de sprekende computer steken vóór de 'kriebel'-vallei. Klabbers: "Het menselijk spraaksignaal kent zoveel nuances, die zijn nog altijd moeilijk om in een systeem te verwerken. Dat blijkt ook wel uit het feit dat gameontwikkelaars onze software wel gebruiken om te kijken hoe hun teksten klinken in het spel, maar uiteindelijk de teksten toch laten inspreken door een professionele stemacteur. In die teksten zitten heel veel verschillende emoties ... We kunnen de stemacteur voorlopig dus echt nog niet vervangen, gelukkig maar."

Slechte bedoelingen

Maar zoals de deepfake van Rutte laat zien, kan een computerstem inmiddels behoorlijk realistisch uitpakken. Daarin schuilt ook een gevaar, erkent Van de Walle. "Je kunt de stem van Biden of Poetin van alles laten zeggen. Ik geloof dat er ook al mensen zijn opgelicht doordat anderen met een gekloonde stem naar de bank belden. Daarom zetten bedrijven vaak een onzichtbare code in hun synthetische stemmen, zodat je altijd kunt aantonen dat deze nep zijn."

Lang niet elke producent zal van dit morele besef doordrongen zijn. "Er zullen altijd mensen zijn die zo'n techniek op een slechte manier willen gebruiken; dat is nauwelijks te voorkomen", beaamt ook de Tilburgse hoogleraar Louwerse. "Gelukkig weten de meeste mensen dat inmiddels en vertrouwen we niet meer zomaar alles wat we zien en horen. Ethische en juridische kwesties zijn zeker belangrijk bij deze techniek, maar ze mogen denk ik de ontwikkeling niet overschaduwen of belemmeren. Natuurlijk mag je iemand niet ongevraagd imiteren, maar er zijn ook tal van nuttige toepassingen waarvoor je deze techniek juist wél wilt doorontwikkelen, zoals in het onderwijs of in de gezondheidszorg. Of al was het maar om systemen te kunnen maken die imitaties van echt kunnen onderscheiden. Als anderen de techniek wél met slechte bedoelingen gebruiken, kun je de methode maar beter begrijpen, zodat je haar kunt herkennen."

En de computer zal daarbij een onmisbaar gereedschap blijken, aldus Van de Walle. "Zelfs als wij een gekloonde stem niet meer als dusdanig herkennen, zal een computer direct doorhebben dat hij nep is, ook zonder zo'n onzichtbare code. Een synthetische stem is namelijk eigenlijk te perfect, te regelmatig. In natuurlijke spraak zit veel meer variatie, je spreekt een zin nooit twee keer op precies dezelfde manier uit. Zelfs als je die variatie in je systeem inbouwt, is het resultaat nog ergens onnatuurlijk. In de perfectie toont zich dus altijd weer de machine." ←

Permentier

Beu

Lang moet ik niet nadenken als u mij vraagt of er ook lelijke woorden bestaan. Ik denk dan aan *beu*. 'Ik ben het beu': dat klinkt als een natte dweil die je op een stenen vloer laat vallen, en het is even inspirerend. Het kan geen toeval zijn dat *beu* een van de zeldzame woorden is die eindigen op *eu*, op de reu en de keu en nog een handvol andere na, die dat overigens wel sneu vinden. Ook de betekenis is geen feest. Je wil niet achter elk hoekje iemand tegenkomen die zegt dat hij dit of dat beu is. Als ik al iets beu ben, zijn het wel mensen die iets beu zijn.

In de zeventiende eeuw zei je 'beu' als het om eten ging. Je at je beu aan spruitjes, bijvoorbeeld, en dan werden die spruitjes zelf ook beu. Maar de verdere oorsprong is in nevelen gehuld. Er bestaat geen geschreven *beu* van voor het jaar 1600. Waren de mensen voor die tijd nooit iets beu?

Beu is vandaag de dag een bijvoeglijk naamwoord, maar het kan zijn carrière begonnen zijn als een los kreetje, een tussenwerpsel zoals *bah* en *boh*. Er bestond overigens in de veertiende eeuw al een woordje *boy*, dat toen al wees op een slecht humeur: 'Het hevet mi boy' ('ik ben ontstemd') zei je baas als je te lang aan de koffiehoek stond te kletsen. En de uitdrukking *hem boy maken* betekende 'zich boos maken'.

Waar *beu* ook goed voor is, is om het geluid van koeien weer te geven in stripverhalen. En dan nog. Ik woon op het platteland en sommige van mijn buurvrouwen zijn koeien. De hele zomer staan die op een dieet van gras, maar ik heb ze nog nooit 'beuend' horen klagen.

Ze zeggen dan liever 'mmmoeoeoehhh!'

Ludo Permentier