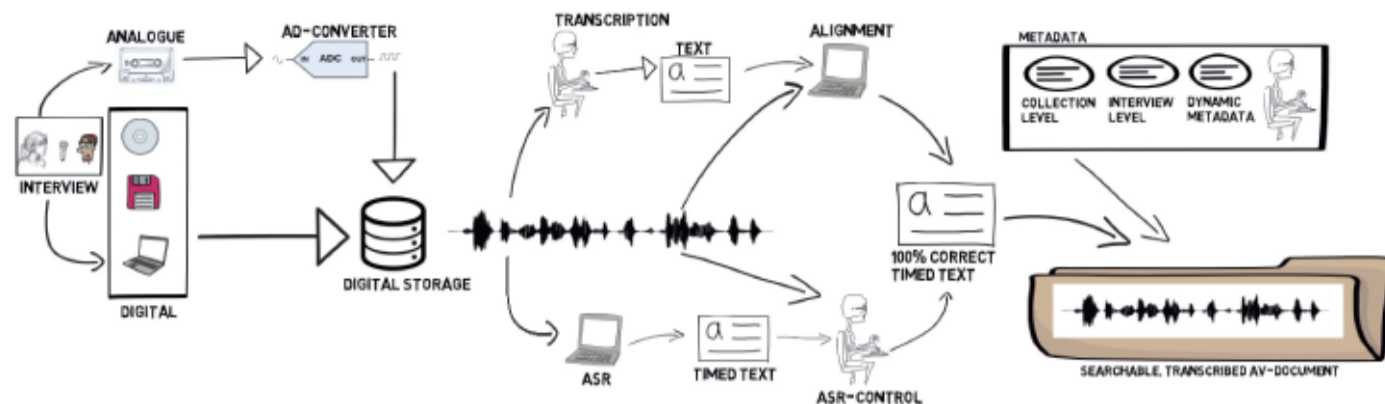


Transcriptie met 95 procent correcte teksten

Spraakherkenning in stroomversnelling

De komst van Deep Neural Networks maakt het analyseren en doorzoeken van spraakopnames een stuk makkelijker. Steeds meer oral history-onderzoekers maken er gebruik van. Erica Renckens



Automatische spraakherkenning (ASR) bestaat inmiddels al een tijdje, maar de resultaten lieten tot voor kort nogal wat te wensen over. Getraind op één gebruiker ging het nog vrij aardig, maar algemene software kon vaak weinig chocola maken van een audiosignaal. Dat maakte het analyseren en doorzoeken van spraakopnames erg lastig. De komst van Deep Neural Networks (DNN), die een beter akoestisch model genereren dan de oude methodes, heeft hier verandering in gebracht.

Tot op de milliseconde

Een DNN-gebaseerde herkenner voor het Nederlands is ontwikkeld door de Universiteit Twente en het

De transcription chain leidt de gebruiker langs de verschillende stappen om zijn analoge opname te verwerken tot een digitaal formaat dat analyse mogelijk maakt. credits Arjan van Hessen

Nederlands Instituut voor Beeld en Geluid. Zij gebruikten hiervoor Kaldi, een gratis open-source toolkit. In samenwerking met de Radboud Universiteit en CLARIN-EU wordt de herkenner opgenomen in een transcription chain: een workflow waarin onderzoekers hun opnamen kunnen digitaliseren, transcriberen, corrigeren en voorzien van metadata. Het resultaat is een reeks woorden die tot op de milliseconde is opgelijnd met de opname.

Taalmodel

“Als de omstandigheden optimaal zijn, is 90 tot 95 procent van de her-

kende tekst nu correct”, aldus Arjan van Hessen, spraaktechnoloog aan de Universiteit Twente. “Veel gebruikers hebben echter geen idee hoe een goede opname te maken. Dikwijls legt men gewoon de smartphone op tafel waardoor ieder contactgeluid keihard in de opname komt. Ook realiseren veel interviewers zich niet dat hun terugkoppeling – ‘ja, ja’, ‘ga door’ – ook in de transcriptie belandt, wat de leesbaarheid niet verhoogt. Liever neem je het gesprek op met aparte headsets.”

Het resultaat van de spraakherkenning hangt niet alleen af van de

kwaliteit van de opname. “Het taalmodel voorspelt de kans op woord X gegeven de woorden Y en Z. Het woordgebruik verschilt per gespreksonderwerp, dus idealiter zijn er ook verschillende taalmodellen. Maar dat is nu nog te arbeidsintensief. We werken aan de mogelijkheid om onderzoekers zelf woordenlijsten te laten toevoegen.”

Toekomstig gebruik

Ook hoopt het team de transcription chain verder uit te breiden met een editor waarmee de output van de spraakherkenner eenvoudig gecorrigeerd kan worden. “Daar zijn

wel al toepassingen voor, maar die zijn niet altijd gebruiksvriendelijk. Of de tijdinformatie gaat verloren waardoor zoeken in het geluidssignaal niet meer mogelijk is.” Van Hessen ziet het gebruik van ASR bij de overheid de laatste tijd sterk toenemen. Zo wordt de herkenner momenteel getest door de politie en de FIOD. “En ook onderzoekers die gebruikmaken van gesproken getuigenissen ontdekken steeds meer het gemak ervan. Met de transcription chain zal dit hopelijk nog verder toenemen.”

oralhistory.eu/workshops/transcription-chain

Hoe meet je Open Science?

Een sticker voor goed gedrag

Rens van de Schoot schreef in het boek *Hoe zwaar is licht. Meer dan 100 dringende vragen aan de wetenschap dat wetenschappers in een glazen huis op een marktplein moeten gaan werken.* Heidi Berkhout

Naar aanleiding van diverse fraudegevallen, meldingen in de media van onverantwoord toepassen van wetenschap, en de replicatiecrisis die diverse vakgebieden boeit, zijn tal van initiatieven ontstaan om de goede naam van de wetenschap te herstellen. Zo verklaarden diverse organisaties de ambities uit het Nationaal Plan Open Science na te streven.

Nationaal Plan

Sander Dekker, demissionair staatssecretaris van Onderwijs, Cultuur en Wetenschap, staat aan de basis van dit plan. Hij roept op tot meer Open Science, waarmee hij bedoelt: (1) het open access publiceren van wetenschappelijke publicaties – dat wil zeggen gratis te lezen; (2) optimaal (her)gebruik van onderzoeksdata – door bijvoorbeeld data FAIR te deponeren en te hergebruiken via archieven, zoals



EASY, het online archiveringssysteem van DANS; (3) het aanpassen van evaluatiesystemen – in het landelijke protocol voor evaluatie van wetenschappelijk onderzoek (het Standard Evaluation Protocol, opgesteld door de KNAW, VSNU en NWO) is bijvoorbeeld het punt ‘productiviteit’ komen te vervallen.

Sticker voor open data

Een belangrijk thema is het beschikbaar maken van onderzoeksdata. “Tegenwoordig krijg je bij sommige tijdschriften een sticker

als je je data openbaar hebt gepubliceerd. Dat is een mooie bijdrage aan Open Science. We moeten echter oppassen dat zo’n sticker zich niet tegen ons keert”, stelt Rens van de Schoot, universitair hoofddocent bij de Universiteit Utrecht, lid van De Jonge Akademie en betrokken bij de ontwikkeling van het Nationaal Plan Open Science. “Het aantal stickers gaan tellen bij beoordelingen lijkt een slim plan. Immers, meer stickers betekent meer Open Science. Maar wat te doen met al die onderzoekers die hun data niet openbaar mogen maken en dus

Van de Schoot adviseert in zijn bijdrage ‘Wat zijn de regels van de wetenschap en liggen die voor altijd vast?’ in het boek ‘Hoe zwaar is licht. Meer dan 100 dringende vragen aan de wetenschap’ dat wetenschappers uit hun ivoren toren moeten komen om de wetenschap opener te maken. “Beter is om de wetenschap in een glazen huis uit te voeren. Hiermee wordt wetenschapsbeoefening zichtbaar voor iedereen, maar we moeten niet doorslaan.” foto Joe Mabel

nooit zo’n sticker kunnen verdienen? Denk aan medische gegevens. Moet je dan alleen maar werken aan projecten waarvan de data openbaar gemaakt kunnen worden?”

Zelf stappen zetten

Gelukkig zijn ook voor dit soort typen data genoeg oplossingen. “Tijdens een bijeenkomst over De Huiskamer van de Wetenschap, een initiatief van De Jonge Akademie (zie kader), noemde Peter Doorn, directeur van DANS, bijvoorbeeld dat data nooit ‘zomaar’ gedeponerd worden. Onderzoekers kunnen zelf

bepalen of data openbaar toegankelijk zijn.” Van de Schoot vervolgt: “Het gaat erom dat we met z’n allen bedenken wat per wetenschappelijk veld nodig is om de volgende stappen te zetten. We hoeven echt niet allemaal in een glazen huis op de markt te gaan werken, maatwerk kan en mag. Maar als we allemaal vandaag nog één stap zetten om een beetje opener te zijn, dan zou dat al een enorme impact hebben.”

Kijk voor alle initiatieven op het gebied van Open Science op openscience.nl

Huiskamer

van de wetenschap

De Jonge Akademie wil weg van een onderzoekscultuur van ‘zo doen we het nu eenmaal hier’ naar een opener discussie over wetenschapsbeoefening. Het daartoe in het leven geroepen project Huiskamer van de Wetenschap bestaat uit rondreizende huiskamerbijeenkomsten en een website (in wording). Zo kunnen jonge wetenschappers kennismaken van discussies over de onderzoeksmores in hun vakgebied en daar zelf over meepreten. Denk aan thema’s als auteurschap, diversiteit en werkdruk.

De Jonge Akademie is onderdeel van de KNAW. dejongeakademie.nl