Universiteit Utrecht

AUTOMATISCHE SPRAAKHERKENNING VOOR INFANT DIRECTED SPEECH

Jolanda Vermeulen
6265723
Begeleider: Frans Adriaans
Tweede beoordelaar: Alexei Nazarov
2 juli 2021
Bachelor kunstmatige intelligentie
Universiteit Utrecht
7,5 ECTS

Abstract:
In dit onderzoek worden drie benaderingen onderzocht om een aanbeveling te doen voor een methode om een automatische spraakherkenning (automatic speech recognition, ASR) voor de manier van praten tegen kinderen. De eerste benadering kijkt naar hoe eerdere ASR gemaakt zijn voor nieuwe talen. Als tweede wordt er gekeken naar hoe er ASR is gemaakt voor dialecten en als derde wordt er gekeken naar de zero-resource methode. De aanbeveling is om twee benaderingen te combineren en met behulp van Kaldi (Povey et al., 2011), SpecAugment (Park et al., 2019) en Sequitur (Bisani & Ney, 2008) een ASR te maken met MFCC (Zheng et al., 2001). Verder moeten alle data die er zijn, worden gebruikt, aangevuld met grafeem naar foneem regels en een lexicon met uitzonderingen.

Sleutelwoorden: Automatische spraakherkenning, babypraat, infant directed speech, Kaldi, SpecAugment, Sequitur

Inhoud

1. Inleiding

Automatische spraakherkenning (*automatic speech recognition*, ASR) wordt veel gebruikt in het dagelijks leven, zoals op onze telefoons. Automatische spraakherkenning heeft verschillende algoritmes. Voor elke taal is er een andere spraakherkenning. Nog niet voor alle talen of alle manieren van praten is er een ASR. Met elke ASR die wordt gemaakt, wordt de techniek beter en kan het zelf leren van de ASR ook worden toegepast op andere technieken. Verder is het ook belangrijk dat alles herkend kan worden, omdat er dan transcripties van kunnen worden gemaakt die van belang zijn bij verschillende onderzoeken over taal. Zo wordt er onderzoek gedaan naar de manier van praten tegen kinderen en hoe dit bijdraagt aan hoe kinderen een taal leren (Cooper & Aslin, 1994). Door middel van een transcriptie kan er makkelijker worden gezocht naar bepaalde target woorden. Deze woorden zijn makkelijker terug te vinden in een transcriptie. Als de ASR goed werkt, kan er zelfs een woord gezocht worden door de ASR zonder dat er een transcriptie nodig is. Een transcriptie met de hand maken kost veel tijd, vooral als er uren aan opnames zijn. Daarom is het gebruik van een ASR als hulpmiddel gewenst. De ASR kan dan de opnames transcriberen en het onderzoek naar taal een stuk vlotter laten verlopen. Een ASR die een standaardtaal kan herkennen, kan bijvoorbeeld deze taal nog niet herkennen als deze door een kind gesproken wordt (Benzeghiba et al., 2007). Een andere taal waar nog geen automatische spraakherkenning voor is, is de manier van spreken tegen baby's en kleine kinderen. Deze manier van spreken heet *infant directed speech* (IDS) (van der Klis et al., 2020). Een reden dat hiervoor een ASR gewenst is, is dat er meer onderzoek gedaan kan worden naar IDS. IDS helpt kinderen met het leren van een taal, omdat het helpt om de aandacht van de kinderen vast te houden. IDS kan ook helpen om een algoritme beter een taal te leren. Een goede start hiervoor is een ASR voor IDS. Er zijn verschillende soorten ASR voor het standaard Engels, maar er is nog geen ASR voor IDS in het Engels. Het gebruik van de ASR voor het standaard Engels voor IDS is niet mogelijk. De ASR maakt dan teveel fouten, omdat de ASR IDS niet goed kan herkennen (van der Klis et al., 2020). Het is dus nodig om een nieuwe ASR te maken voor IDS. De belangrijkste vraag in dit onderzoek zal zijn hoe er een automatische spraakherkenning gemaakt kan worden voor de manier van spreken naar kinderen. Er zijn verschillende manieren om een ASR te maken. Er kan een nieuwe ASR gemaakt worden, er kan een bestaande ASR worden aangepast of er kan een nieuwe ASR gemaakt worden door middel van de zero-resource methode. In dit onderzoek worden deze drie benaderingen onderzocht om erachter te komen welke benadering geschikt is om een ASR te maken voor IDS. Deze beoordeling wordt gebaseerd op de fouten die een ASR maakt, de *word error rate* (WER) en hoeveel data er nodig zijn om deze ASR te maken. Een lage WER betekent dat de ASR weinig fouten maakt, dit wordt in een percentage weergegeven. Een goede WER is maximaal 20%. De hoeveelheid data is belangrijk, zodat het vergeleken kan worden met hoeveel er is van IDS en of het de moeite waard is om meer data te verzamelen voor IDS voor een ASR.

De relevantie van dit onderzoek voor kunstmatige intelligentie is de automatische spraakherkenning die er moet worden gemaakt of worden aangepast. Dit is een automatisch systeem dat zelf leert om nieuwe data te herkennen en dit om te zetten naar een transcriptie. Ongecontroleerd leren is een belangrijk onderdeel van kunstmatige intelligentie, in dit onderzoek is het terug te zien in het zelf leren herkennen van een taal. In dit specifieke onderzoek is het belangrijk dat de ASR nieuwe data kan transcriberen. Er is veel onderzoek naar IDS en er zijn vele uren aan geluidsopnames beschikbaar. Om dit allemaal met de hand te moeten transcriberen zou veel te lang duren. Het is dus belangrijk dat een ASR de geluidsopnames van IDS kan transcriberen, zodat onderzoek naar IDS een stuk vlotter kan verlopen.

De data die er zijn met transcriptie van IDS, zijn beperkt. Er zijn onder andere 203 audio opnames van 75 minuten. Deze 75 minuten bestaan niet uit continue spraak. Van de 203 audio opnames zijn er 35 zonder transcriptie. Deze audio opnames zijn in het Amerikaans Engels. Deze audio opnames zijn te vinden in de Brent corpus op CHILDES (Brent & Siskind, 2001; Macwhinney, 2000). Er is ook 14 uur aan getranscribeerde audio van IDS in het Japans (Reiko et al., 2006). In dit onderzoek wordt gefocust op wat er bestaat voor het Engels, omdat IDS in het Japans anders is. Met deze data kan een ASR worden getraind. In dit onderzoek wordt gekeken naar wat IDS is en welke eigenschappen IDS heeft om erachter te komen waar op gelet moet worden in een ASR. Ook wordt er gekeken hoe IDS verschilt van ADS (de Boer & Kuhl, 2003). Een groot verschil tussen IDS en ADS is de hoge pitch (Adriaans & Swingley, 2017). Een ASR die getraind is op ADS kan slechter spraak van kinderen herkennen door de hogere pitch (Ghai & Sinha, 2015). IDS heeft ook deze hogere pitch, dus de ASR moet de hoge pitch wel goed kunnen herkennen. In dit onderzoek wordt er bij de eerste benadering gekeken hoe een nieuwe ASR voor een taal wordt gemaakt. Hierbij wordt gelet op de hoeveelheid en het soort data die ze gebruiken. Ook wordt erop gelet wat voor modellen er worden gebruikt en wat voor bestaande tools er al zijn. Een voorbeeld van een nieuwe IDS die is gemaakt voor het Duits, is de ASR van Milde en Koehn (2018). Zij zagen dat er wel verschillende ASR van hoge kwaliteit zijn voor het Engels die gebruik maken van de Kaldi toolkit, maar nog niet voor het Duits. Uiteindelijk hebben ze een vrij toegankelijk ASR-model gemaakt voor het Duits met behulp van de Kaldi toolkit (Milde & Koehn, 2018). De Kaldi toolkit is een open source toolkit waarbij verschillende algoritmes worden gebruikt, waarmee er een ASR kan worden gemaakt (Povey et al., 2011). Vervolgens wordt gekeken naar de motieven waarom ze bepaalde modellen hebben gebruikt. Voor de tweede benadering wordt er gekeken naar de manier waarop een ASR wordt gemaakt van een dialect in een taal waar al een ASR voor bestaat. Er wordt onderzocht wat ze aanpassen of wat ze anders doen als er een nieuwe ASR wordt gemaakt voor een dialect. Ook wordt hier gekeken naar de gebruikte data. Biadsy (2011) heeft een ASR verbeterd door er een herkenning van dialecten aan toe te voegen. Er wordt onderzocht wat voor manieren er nog meer bestaan en hoe dit is aangepakt. Tenslotte wordt er onderzoek gedaan naar de derde benadering. Dit is de zero-resource methode. Hier wordt gekeken naar ASR die gemaakt is met weinig tot geen data. Zero-resource spraaktechnologie werkt zonder linguïstieke kennis. Al deze kennis moet het systeem zelf verkrijgen door middel van de audio die het krijgt (Jansen et al., 2013). Met de zero-resource methode zal de ASR de taal helemaal zelf moeten leren herkennen. Deze drie benaderingen vormen de basis voor de deelvragen. Voor deze deelvragen worden alle voor- en nadelen benoemd en wordt beschreven wat ervoor nodig is voor deze benaderingen. Met deze informatie wordt er een aanbeveling voor de beste aanpak om een automatische spraakherkenning te maken voor de manier van spreken tegen kinderen.

De eerste paragraaf zal gaan over IDS. Hier wordt toegelicht wat IDS inhoudt en hoe het verschilt van *adult directed speech* ofwel ADS, het praten tegen een andere volwassene. De tweede paragraaf zal de eerste deelvraag gaan beantwoorden: Kan er een automatische spraakherkenning worden gemaakt voor IDS door het te zien als een nieuwe taal? Eerst wordt er gekeken naar hoe er eerder een ASR voor een nieuwe taal gemaakt is. Hierna wordt er beschreven hoe dit toepasbaar is op IDS. De volgende twee paragrafen volgen dezelfde opbouw. De derde paragraaf zal focussen op hoe een ASR word gemaakt als IDS wordt gezien als dialect en de vierde paragraaf zal beschrijven hoe het mogelijk is om een ASR te maken met de zero resource methode. Vervolgens wordt dit afgesloten met de aanbeveling hoe er een ASR gemaakt kan worden voor IDS.

2. Wat is Infant Directed Speech (IDS)?

De taal waarmee men naar kinderen praat, is anders dan de taal waarmee men naar andere volwassenen praat. Dit wordt *infant directed speech* (IDS) genoemd of ook wel babypraat. De manier waarop moeders tegen hun kinderen praten, is een bijzondere vorm van taal. Het verschilt veel van de normale manier van praten. In deze paragraaf worden de grote verschillen van het praten tegen een kind en het praten tegen een volwassene uitgelicht.

De reden dat moeders op een andere manier praten tegen een kind dan tegen een volwassene is dat de aandacht van het kind op deze manier er beter bij wordt gehouden en het kind ook werkelijk luistert. Verder is deze manier van spreken ook een bron van taalkundige informatie voor het kind in de periode dat het kind het snelste taal leert. Onderzoek laat zien dat kinderen weinig of niet naar gesprekken luisteren tussen volwassenen, maar wel als er sprake is van IDS (Fernald & Simon, 1984).

Het belangrijkste is dat de toon een stuk hoger ligt als er tegen een kind wordt gepraat in plaats van tegen een volwassene. Dit gebeurt om de aandacht van een kind te bewaren (Fernald, 1985). Er is ook sprake van een hyperarticulatie. In IDS worden hoekpunt klinkers hypergearticuleerd in vergelijking met ADS. Hoekpunt klinkers zijn de klinkers [ie], [oe] en [aa]. Dit worden hoekpunt klinkers genoemd, omdat je met alle klinkers een klinkerdriehoek kunt maken. Deze klinkers zijn dan de hoeken van de driehoek. Elke klinker die uit te spreken is, valt in deze driehoek en elke andere klinker dan de hoekpuntklinkers valt ook binnen deze punten. Aangezien de hoekpuntklinkers in IDS worden hypergearticuleerd, staan de klinkers verder van elkaar weg. Het verschil in de uitspraak van de klinkers wordt hierdoor groter. Toch zijn er ook specifieke contrasten die niet zijn verhoogd. IDS is niet met alle tekens consistent duidelijker (Adriaans & Swingley, 2017). Het is dus belangrijk dat de ASR verschillende uitspraken kan herkennen.

We zien dat de grootste verschillen tussen IDS en ADS zijn dat er met een hogere toon wordt gepraat, een meer variërende toon, langere pauzes en kortere spreekeenheden (Piazza et al., 2017). Dit zijn de belangrijkste variaties op ADS die moeten worden onthouden voor het maken van een spraakherkenning. Er zijn 203 keer 75 minuten aan opnames met spraak erin, waarvan 35 zonder transcriptie. Dit betekent dat er 210 uur aan opnames met spraak erin is die een transcriptie hebben.

Deze verschillen kunnen de oorzaak zijn van het feit dat een ASR voor de standaardtaal geen IDS kan herkennen. Dit is gezien bij van der Klis et al. (2020). Hier werd een ASR voor het Nederlands gebruikt voor IDS in het Nederlands. De ASR deed het duidelijk minder goed in het herkennen van IDS dan ADS. Ghai en Sinha (2015) hebben de ASR verbeterd, omdat deze de spraak van kinderen niet kon herkennen vanwege de hoge toon. Hier hebben ze een algoritme toegevoegd die de hoge toon verschillen normaliseert door verbeterende *smoothing* van de toon harmonica in het spraakspectrum.

3. Benadering 1: IDS als nieuwe taal

In deze paragraaf ga wordt er gekeken naar hoe er een ASR gemaakt kan worden voor een nieuwe taal. Hiervoor worden er verschillende technieken bekeken en ASR die deze technieken hebben gebruikt. Er zijn al veel soorten ASR gemaakt, hierdoor verbeteren de technieken voor ASR ook. Zo zijn er ook technieken en toolkits die kunnen worden gebruikt om makkelijker een ASR te maken. Zo is er SpecAugment (Park et al., 2019) om data te vergroten en is Kaldi op het moment de meest gebruikte toolkit. Eerst zullen deze twee programma's worden uitgelegd. Vervolgens worden er verschillende soorten ASR laten zien die

gebruik maken van deze programma's. SpecAugment is een techniek die veel wordt gebruikt om data te vergroten. Deze techniek past de data aan, zodat het lijkt alsof er meer data zijn. SpecAugment bestaat uit drie soorten vervormingen van het mel spectrogram. De eerste zorgt ervoor dat tijdreeks vervormd wordt in de tijdrichting. De tweede- en derde vervorming zijn tijd- en frequentiemaskering. Hier maskeren ze een blok van opeenvolgende tijdstappen of mel frequentie kanalen. SpecAugment zorgt ervoor dat een ASR van een *overfitting* probleem wordt veranderd naar een *underfitting* probleem (Park et al., 2019; Yu et al., 2020). Deze techniek is handig voor IDS, omdat er weinig data is. Zo kan de hoeveelheid data vergroot worden en het model getraind worden met meer data.

Een veel gebruikte toolkit in het maken van ASR is de Kaldi toolkit. Een groot voordeel van Kaldi tegenover de toolkits zoals HTK (Young, 1997) en Sphinx (Lee et al., 1990) is dat er nog actief wordt bijgedragen aan Kaldi en de geavanceerde *state of art* technieken van Kaldi (Guglani & Mishra, 2018). Het doel van deze toolkit was om een flexibele code te maken. Deze code is makkelijk te begrijpen, aan te passen en uit te breiden. Met deze toolkit proberen Povey et al. (2011) het makkelijker te maken om een ASR te maken. De toolkit gebruikt het Gaussian Mixture Model (GMM) om een taal te leren. Met Kaldi is het mogelijk om een accurate ASR te maken. De doelgroep van Kaldi bestaat uit wetenschappelijke onderzoekers (Guglani & Mishra, 2018). Voor deze toolkit is geen specifieke hoeveelheid aan data nodig, maar er zal worden genoemd hoeveel data elk onderzoek gebruikt die deze toolkit heeft gebruikt.

De Kaldi toolkit is bijvoorbeeld gebruikt door Milde en Koehn (2018) zoals eerder is genoemd. Zij zagen dat er hoge kwaliteit ASR is met behulp van Kaldi voor het Engels, maar niet voor het Duits. Het doel was om een soort gelijk toegankelijk systeem met uitleg te maken voor het Duits met Kaldi, zoals deze bestaat voor het Engels. Ze gebruiken het *Spoken Wikipedia Corpus* (Wikipedia contributors, 2021) en Tuda-De (Radeck-Arneth et al., 2015) die meer soorten tekstgenres bevatten. Dit gaf ze 268 uur aan conservatieve *pruning* en 412 uur aan minimale *pruning*. Al deze data zijn getranscribeerd. Later komen ze erachter dat in de Tuda-DE data een paar foute transcripties zitten. Nadat ze deze hadden verwijderd, bleef er 250 uur over van de 268 uur en slechts 375 uur van de 412 uur. Ze gebruiken Sequitur (Bisani & Ney, 2008) als programma om te trainen van grafeem naar foneem (G2P). Vervolgens trainen ze 3-gram en 4-gram taal modellen met Kneser-Ney smoothing (Kneser & Ney, 1995). Verder gebruiken ze het stappenplan van Kaldi en maken ze een vrij toegankelijke ASR voor het Duits die goed aan te passen is voor andere talen zoals het Nederlands. De *word error rate* (WER) is 14,4% voor TDNN-HMM (*Time-Delayed Neural Networks-Hidden Markov Model*). Voor deze methode zouden de data die nu bestaan voor IDS verdubbeld moeten worden. De data verdubbelen is nog mogelijk.

Kaldi is een veelbelovende kit en daarom wordt er in dit onderzoek nog gekeken naar twee andere ASR door middel van Kaldi. Dit is een ASR voor het Punjabi (Guglani & Mishra, 2018) en een ASR voor het fluisteren in het Pools. In het onderzoek voor een ASR voor het Punjabi vergelijken ze twee compacte representaties. De eerste is een set van mel-frequentie cepstrum coëfficiënten (MFCC) (Zheng et al., 2001) en de tweede door middel van perceptuele lineaire voorspellende techniek (PLP) (Hermansky, 1990). Om deze ASR te maken hebben ze gebruik gemaakt van 240 uur spraak met transcriptie. Het onderzoek laat zien dat de MFCC een lager aantal fouten had dan de PLP. Uiteindelijk kozen ze voor de akoestische modellering (tri2b_bmmi) discriminerend getraind bij verhoogde maximale wederzijdse informatie (bMMI) met MFCC. LDA en MLLT zijn gebruikt voor de voorbewerking, omdat informele experimenten hebben laten zien dat decoderen met *minimal phone error* (MPE) akoestisch model iets langer duurt om te verwerken. Zover Guglani en Mishra weten is dit de eerste ASR die voor Punjabi is gemaakt met behulp van Kaldi. De

WER voor deze ASR met MFCC en 2-gram is 21,8% en met de 3-gram is het 21,2%. Deze WER is stuk hoger dan de ASR voor het Duits, maar hier is veel minder data voor gebruikt. Dit maakt deze methode interessant voor het IDS.

Een andere ASR die gemaakt is met Kaldi, is die voor het fluisteren in het Pools. Kozierski et al. (2016) gebruikten Kaldi met SRILM (Stolcke, 2002) om de toolkit uit te breiden. Daarna gebruikten ze nog Sequitur als programma om de conversie te maken van grafemen naar fonemen (G2P). Voor het corpus die ze gebruikten, hadden ze fluistertaal en normale taal. De fluistertaal en de normale taal bestonden allebei uit ongeveer 9 uur aan materiaal met transcripties. Deze hebben veel gelijke kenmerken. Dit is te zien, omdat er een verbetering is, nadat ze de test data hebben toegevoegd. Ze trekken de conclusie dat ze een groter corpus nodig hebben, omdat er een te groot verschil is tussen de trainingsdata en testdata. De WER is door de verschillen ook niet duidelijk.

Kaldi is een toolkit waarmee een ASR voor IDS zou kunnen worden gemaakt. Het eerste onderzoek gebruikt totaal 600 uur aan data en deze heeft een heel lage WER. Het tweede onderzoek gebruikt 240 uur aan data en de WER is wel hoger dan het eerste onderzoek, maar nog steeds rond de 20%. Het laatste onderzoek maakt gebruik van 18 uur aan data en heeft de normale spraak met fluisteren door elkaar. Deze ASR werkt ook niet goed genoeg. Om gebruik te maken van Kaldi is het dus nodig om minstens 240 uur aan getranscribeerde spraak te hebben, maar het is duidelijk dat data zorgt voor een lagere WER. Van IDS is er 210 uur aan data. Om een goede ASR voor IDS te kunnen maken met behulp van de Kaldi toolkit is er meer data nodig.

Zoals Kaldi zijn er ook andere technieken, zoals wordt gedaan voor een ASR voor het Mandarijn. Zhang en Liu (2020) hadden als doel om een betere ASR te maken dan de ASR die al bestaat voor het Mandarijn. Hiervoor gebruiken ze het open-source Mandarijn corpus Aishell-1 en een extra corpus. Totaal hebben ze 150 uur aan spraak met transcriptie. Ze gebruiken Kaldi (Povey et al., 2011) om kenmerken te vinden en ESPnet om het model te trainen. ESPnet is een open source platform voor *end-to-end* spraak verwerking. Het is geïntroduceerd door Watanabe et al. (2018). ESPnet heeft een enkel neuraal netwerk architectuur. Voor het deep learning in het neurale netwerk wordt gebruik gemaakt van Chainer (Tokui et al., 2015) en PyTorch (Paszke et al., 2017). Dit versimpelt de training en herkenning van de hele ASR. Verder gebruikt ESPnet dezelfde stijl als de Kaldi toolkit voor dataverwerking, namelijk kenmerkherkenning en stappen voor de set-up voor een spraakherkenning. Voor de ASR voor het Mandarijn trainen Zhang en Liu een long short-term memory (LSTM) taal model (Hori et al., 2017; Brownlee, 2020) met het extra corpus en daarna vervangen ze de Bi-LSTM encoder met de simple recurrent units (Bi-SRU). Hierna gebruiken ze regularisatie methodes en op SpecAugment gebaseerde data. Dit model is de beste die ze hebben kunnen maken en is sneller voor trainen en decoderen. Er wordt in het onderzoek genoemd dat deze keuze het laagste percentage aan fouten geeft, er wordt echter geen percentage genoemd. Zhang en Liu zullen verder gaan met het vervangen van LSTM gebaseerde decodeerders met SRU gebaseerde decodeerders. Ze laten hier zien dat het mogelijk is om Kaldi te gebruiken met minder data. Het nadeel is dat er geen exacte vergelijking is hoe goed de ASR is, omdat er geen WER is gegeven.

Er zijn nog verschillende andere toolkits en programma's die bestaan om te helpen met het maken van een ASR, maar de meest gebruikte is Kaldi. De andere mogelijkheden zal worden nog even kort noemen. HTK is (Young, 1997) een Hidden Markov Model Toolkit. Deze toolkit kan voor alles worden gebruikt, maar wordt het meest gebruikt voor spraakherkenningen. Ze hebben verschillende modules en tools beschikbaar

gemaakt om het werken met HMM zo makkelijk mogelijk te maken (*HTK Speech Recognition Toolkit*, 2016). Dan is er Sphinx die gebruik maakt van HMM's met LPC afgeleide parameters (Lee et al., 1990). Sphinx wordt niet veel meer gebruikt sinds Kaldi en Pytorch (*Open Source Speech Recognition Toolkit*, 2019) zijn gekomen. PyTorch is een veel gebruikt framework voor deep learning. PyTorch maakt het mogelijk om dynamisch functies te maken en het verloop van deze functies te bekijken. Het maakt het makkelijk om met nieuwe deep learning architecturen te experimenteren (Ketkar, 2017). Verder bestaan nog Julius (Lee et al., 2001), RWTH (Rybach et al., 2009) en Returnn (Doetsch et al., 2017). Deze worden vrijwel niet gebruikt. PyTorch is een goed framework, maar voor ASR gaat de voorkeur toch naar Kaldi.

De meeste ASR die tegenwoordig wordt gemaakt voor talen waar data beschikbaar voor zijn, maken gebruik van Kaldi. De andere technieken zijn ouder en minder mee ontwikkeld door nieuwe technieken of zijn veel lastiger in gebruik. Als IDS wordt gezien als een taal waar een nieuwe ASR voor gemaakt moet worden dan is Kaldi de toolkit om te gaan gebruiken. Er is eerder gezien dat hoe meer data er beschikbaar waren hoe lager de WER werd. Een minimaal aantal van 240 uur is gewenst, maar 600 uur is nog beter. Dit is een nadeel voor deze methode, want er is op het moment 210 uur aan data voor IDS. Als alleen dit wordt gebruikt, zal de WER nog erg hoog zitten. Voor deze methode zal er dus meer uren aan data moeten worden getranscribeerd om een goede ASR te kunnen trainen. Voordeel is dat de data kan worden vergroot met behulp van SpecAugment. Het belangrijkste voordeel van het gebruik van Kaldi is dat het makkelijk in gebruik is. Ze hebben stappenplannen om te gebruiken en er zijn al veel anderen geweest die het succesvol hebben gebruikt. Deze methode is dus een geschikte mogelijkheid voor het maken van een ASR voor IDS.

4. Benadering 2: IDS als dialect

In deze benadering wordt er gekeken naar IDS als dialect. IDS is geen dialect, maar een variatie op de taal. De meeste woorden en klanken zijn gelijk met de standaardtaal. De uitspraak van sommige woorden is anders dan een standaardtaal en soms is de woordvolgorde anders. Er kan in de derde persoon worden gepraat of er kunnen kortere zinnen worden gebruikt. In de woordenschat is er weinig verschil in IDS en de standaardtaal. Het is misschien niet nodig om een nieuwe ASR te maken, maar mogelijk om een bestaande aan te passen, zoals voor een dialect kan worden gedaan. Daarom wordt in deze benadering gekeken naar IDS als een dialect van de standaardtaal.

In dialecten zijn er verschillende fonetische en taalkundige verschillen die je niet kunt onderscheiden door middel van duidelijke grenzen. Toch zijn er drie duidelijke verschillen met standaardtaal: de uitspraak, de woordenschat en de woordvolgorde (Hirayama et al., 2015). Door deze drie punten te bekijken kan er een ASR worden gemaakt voor het dialect. Hirayama et al. onderzochten een ASR die verschillende dialecten moest herkennen. Hirayama et al. gaan ervan uit dat de woordvolgorde niet verandert en ze focussen op de uitspraak en woordenschat. Om deze ASR te maken, gebruiken ze bestaande ASR's die een dialect kunnen herkennen en combineren. De methode van Hirayama et al. is niet toe te passen op een ASR voor IDS, want er zijn nog geen modellen voor het herkennen van IDS. Zodra er verschillende soorten ASR komen voor IDS kan deze methode worden toegepast. Dan zouden er verschillende soorten ASR voor IDS kunnen worden gebruikt om alle soorten van IDS te kunnen herkennen.

Masmoudi et al. (2017) maken een ASR voor het Tunesische dialect. De taal die zij als standaardtaal gebruiken, is het Arabisch. Voor Arabisch bestaan verschillende soorten ASR, maar deze kunnen niet het Tunesische dialect herkennen. Masmoudi et al. hebben eerst gekeken naar de karakteristieken van het

Tunesische dialect. Vervolgens gebruiken zij een grafeem naar foneem (G2P) conventie om een set van G2P regels en een lexicon van uitzonderingen te maken. Verder maken ze een corpus genaamd TARIC met het *Tunisian Railway Transport Network domain*. Dit hebben ze handmatig getranscribeerd en bestaat uiteindelijk uit 10 uur getranscribeerde spraak en nog 10 uur spraak zonder transcriptie. Met dit corpus en de G2P regels kunnen ze de Kaldi toolkit gebruiken van Povey et al. (2011). Hiermee maken ze een ASR voor het Tunesische dialect. Deze ASR heeft uiteindelijk een WER van 22,6%. Het verschil met deze methode met Kaldi om een dialect te herkennen en de vorige methodes met Kaldi om een ASR voor een nieuwe taal te maken, is dat bij deze methode om een dialect te herkennen het model ook G2P regels mee krijgt en een lexicon met uitzonderingen. Zo kan het model leren met minder data. Het voordeel hiervan is dat er geen extra spraak meer hoeft te worden getranscribeerd. Het nadeel van deze methode is dat de regels en het lexicon nog moeten worden gemaakt.

In Chinese dialecten wordt er veel gewisseld tussen talen in een zin. Lyu et al. (2006) noemen uit een artikel van Chen (2004) dat Mandarijn en Thais vaak door elkaar worden gesproken in alledaagse gesprekken. De standaard ASR kunnen dit niet goed herkennen. Het doel van Lyu et al. is om een ASR te maken die dit wel kan herkennen. Ze maken gebruik van een one-pass kader waar gebruik wordt gemaakt van slechts één stadium in plaats van drie stadia, zoals in een multi-pass kader. Een multi-pass kader splitst de herkenningsopdracht in het deel waar de talen worden gesplitst en dan de identificatie van die talen. Bij een one-pass kader wordt het herkend in één stadium. Met een corpus van 11,3 uur in het Mandarijn, 11,2 uur in het Thais en 4,41 uur in code-switching tussen Mandarijn en Thais, hebben ze deze twee methodes vergeleken. Het verschil was niet groot tussen de twee methodes, maar het one-pass kader kan de spraak herkennen zonder taal identificatie. De ASR, getest op 20.000 woorden, had een WER van 20,02%. Voordeel van deze methode is de lage WER op een klein corpus waarop getraind is. Het nadeel is dat deze methode gefocust is op het herkennen van twee talen. IDS is één taal. Het is dus lastig om te vergelijken of deze methode het ook goed zou doen op IDS.

Een ASR maken voor IDS, zoals er ASR wordt gemaakt voor dialecten, is mogelijk. De verschillende soorten ASR die hier voor de dialecten worden gemaakt, maken een totaal nieuwe ASR. Ze passen geen bestaande ASR aan van de standaardtaal om een ASR te maken die het dialect kan herkennen. In deze sectie is de methode van het Tunesische dialect het meest geschikt voor het IDS. Bij het Mandarijn en Thais ligt de focus op het herkennen van twee talen. Dit is voor IDS niet nodig. De methode van het Tunesische dialect maakt gebruik van weinig data en heeft een lage WER. Voordeel van deze methode is dat er geen spraak met de hand hoeft te worden getranscribeerd en gebruik kan worden gemaakt van het stappenplan van Kaldi. Wel moeten er nog G2P regels worden gemaakt en een lexicon met uitzonderingen.

5. Benadering 3: IDS als zero-resource

In deze paragraaf wordt er gekeken naar de zero-resource methode. De meeste soorten ASR die nu bestaan, zijn gemaakt door veel gelabelde data. De taalmodellen worden getraind met deze grote hoeveelheden data (Versteegh et al., 2016). Toch zijn er veel talen waar weinig over bekend is. Voor deze talen is het ideaal om een ASR te hebben die het kan transcriberen. Zo kan de taal sneller onderzocht worden en kan er meer worden geleerd over de taal. Met zero-resource kan je een ASR maken van een taal waar men niks van weet. Zero resource is zonder getranscribeerde data (Versteegh et al., 2016). Zero resource opereert namelijk zonder taalkundige kennis waar een standaard ASR op werkt (Jansen et al., 2013). Denk hier aan

een deel getranscribeerde spraak, taal modellen en uitspraak. De bedoeling is dat een zero resource dit zelf leert. Het idee hiervan is gebaseerd op hoe baby's taal leren zonder voorkennis.

Zero-resource is nog een methode die weinig wordt gebruikt voor ASR. De WER van deze systemen is nog te laag. Om sneller meer kennis te verwerven, zijn er uitdagingen bedacht waar mensen zonder getranscribeerde data moet gaan werken. In 2015 was de eerste *Zero Resource Speech Challenge.* Versteegh et al. (2016) beschrijven de verschillende aanpakken en resultaten. Er waren twee tracks. Bij de eerste track was het de bedoeling om een kenmerk representatie te maken van gelabelde spraak die het beste de verschillen van fonemen ziet. Bij de tweede track was het de bedoeling dat het systeem als input onaangeraakte spraak bestanden kreeg. Vervolgens moest het systeem als output de meest terugkerende spraakfragmenten teruggeven. Track 1 moet minimaal zo goed zijn als MFCC. Voor deze track bestaat de topline uit posterior grams afgeleid van een Kaldi GMM-HMM systeem met driefoon staten en een bigram woord model. Track twee moet minimaal zo goed zijn als een eerder gemaakte *spoken term discovery system.* Voor de topline hebben ze patronen bestudeerd, die ontdekt zijn bij een adapter grammatica 19 systeem gebaseerd op de foneem annotatie. Er waren vijf papers voor track 1 en twee papers voor track 2. Voor de eerste track, waar de systemen ongecontroleerde topdown informatie op woord niveau hebben gebruikt, hebben ze bij de systemen een nieuwe manier geïntroduceerd om supervisie te gebruiken in een akoestische modellering. De systemen die arbitraire informatie gebruiken, laten een veelbelovende en intrigerende manier zien om dit type informatie te extraheren en te exploiteren. Voor de tweede track introduceren de systemen een ongecontroleerde lettergreepsegmentatie als een stap in gesproken term ontdekking. De verkenning van deze cluster algoritmes biedt een onderbouwing voor deze belangrijke stap en benadrukt de noodzaak van een goede representatie van invoerfuncties.

Deze challenge was er ook in 2017 (Dunbar et al., 2017). In 2017 hebben ze dezelfde tracks als in de 2015 challenge. Voor de eerste track waren er 14 systemen ingezonden en voor de tweede track waren er drie systemen ingezonden. Het was duidelijk dat deze challenge voortbouwde op de vorige challenge. Er is een nieuwe strategie geïntroduceerd: meertalige training. Deze strategie is gebaseerd op de situatie van een meertalig kind. Hier is verdere training met gebruik van neurale netwerken voor gebruikt en de applicatie van tijd series modellen. Het simpele bottom-up clusteren lijkt nog het beste te werken in een ongecontroleerde situatie. Er is nog verder onderzoek nodig om alle werken systematisch te kunnen evalueren. Deze methodes zijn interessant om verder mee te leren, voor de ASR nu is de WER nog te hoog om mee verder te gaan.

Voor het maken van een ASR laten Jansen et al. (2013) zien dat er eerst een keuze kan worden gemaakt of er gesplitst wordt op de woorden of zelfs op de tekens. Normaal is er voor het trainen van een ASR veel uur aan getranscribeerde spraak audio nodig. Als er weinig bronnen zijn voor audio, is het doel om de ASR te verbeteren met minimale kosten. In de workshop (Jansen et al., 2013) zijn er twee manieren bekeken. Bij de eerste is er een meertalig corpus gebruikt om een herkenner te trainen. Dit corpus bestond uit 31 uur spraak in het Duits en Spaans. Ook zat er één uur in het Engels bij. De output van een interne *deep neural network* (DNN) laag is gebruikt om bottleneck kenmerken te maken. Met deze kenmerken werd een Engelse herkenner getraind met grote woordenschat en hetzelfde uur aan getranscribeerde Engelse spraak. Bij de tweede manier hebben ze niet getranscribeerde spraak automatisch laten transcriberen en dat gebruikt om de gelabelde data te vergroten voor de training. Dit heet zelf-training. Hier hebben ze de herkenner die ontstaan was in de eerste manier met het enkele uur aan Engelse spraak, gebruikt om een toegevoegde 14 uur aan Engelse spraak te transcriberen. Het voordeel van deze methode is dat er weinig spraak nodig is

om de ASR te trainen. Er zouden zelfs verschillende talen kunnen worden gebruikt om de ASR te trainen. Het nadeel van deze methode is dat de ASR ongeveer de helft van de tijd fout zit met het herkennen van woorden. Deze methode heeft nog verbeteringen nodig, voordat deze een goede ASR kan vervangen. Er zou bijvoorbeeld nog gekeken kunnen worden naar de tweede manier met een deel getranscribeerde data.

Wiesner et al. (2018) maken ook een ASR voor een taal met weinig data. DARPA's LORELEI programma is het DARPA's *Low Resource Languages for Emergent Incedents* programma. Het doel van DARPA's LORELEI programma is om snel technologie voor talen te ontwikkelen, met name voor talen met weinig middelen. Het is vooral gericht om beter hulp te kunnen bieden met bijvoorbeeld rampenbestrijding (Wiesner et al., 2018). Het doel is om met weinig data een situatie frame (SF) type herkenner te maken. Deze moet drie elementen herkennen in een spraakdocument, namelijk relevantie, situatie type en locatie. Met een inheemse informant hebben ze 15 tot 30 minuten spraak met transcriptie en een paar honderd SF type labels. Uiteindelijk hebben ze een systeem gemaakt die de taal van de plek van het incident kan vertalen naar het Engels. Ook kan het systeem het incident classificeren. Deze combinatie heeft geen transcripties nodig. Ze gebruiken uiteindelijk dus geen ASR om tekst te transcriberen, maar om situaties te herkennen. Voor het doel van dit onderzoek, is deze methode niet geschikt.

Ren et al. (2019) pakken het anders aan. Zij willen tekst naar spraak (TTS) combineren met een ASR. Ze maken een ongecontroleerde methode voor beide. Ze gebruiken de *denoising auto encoder* (DAE) (Vincent et al., 2008). Duale transformatie is wat ze gebruiken om te trainen. Ze transformeren een spraakfragment naar tekst en die gebruiken ze om de TTS te trainen. Vervolgens gebruiken ze het paar van spraakfragment en tekst om de ASR te trainen. Deze manier zorgt ervoor dat het niet goed getraind wordt. Om dit op te lossen gebruiken ze bidirectionele sequentie modelering (BSM). Ze gebruiken Transformer als basismodel. Transformer is het eerste sequentietransductiemodel dat volledig is gebaseerd op aandacht (Vaswani et al., 2017). Transformer keurt voornamelijk het zelf-attentie mechanisme goed. Welke bestaat uit *een multi-head* attentie om cross-positie informatie aan te trekken en een feed forward netwerk om verzekerd te zijn van een niet lineaire transformatie in elke positie. Elke positie wordt gevolgd door resterende connecties en een laag normalisatie. Om het model te trainen gebruiken ze 200 paren van spraak en transcriptie erbij. Dit is ongeveer gelijk aan 20 minuten spraak. Deze paren in combinatie met de DAE, duale transformatie en de BSM geeft de laagste error. De error meten ze op fonemen. De ASR heeft een foneem error rate van 11,7%, maar het is niet bekend wat de WER is. In de toekomst willen ze een model maken dat werkt zonder te trainen op data die al getranscribeerd zijn. Deze methode heeft als voordeel dat er heel weinig paren nodig zijn om het model te trainen. Het nadeel is dat het niet duidelijk is hoe goed dit model werkt op spraak en hoe goed dit model spraak ook kan herkennen en transcriberen.

In een vervolgonderzoek is hierop doorgegaan en is LRSpeech gemaakt (Xu et al., 2020). Xu et al. (2020) stellen voor om modellen van tevoren te trainen en fine-tuning, duale transformatie en kennisdestillatie in LRSpeech om met minder gegevens de nauwkeurigheid van TTS- en ASR-modellen te verbeteren. De modellen worden vooraf getraind op talen waar veel data van zijn. Hier zitten veel paren in van tekst en spraak. Dan gaan ze de TTS en ASR finetunen op de talen met weinig data. Verder wordt er ook weer gebruik gemaakt van duale transformatie om de nauwkeurigheid te verbeteren van TTS en ASR met ongepaarde spraak en tekst. Het verschil hier met het vorige onderzoek is dat ze hier multi-speaker TTS en ASR gebruiken. Voor de transformatie wordt de Transformer gebruikt (Vaswani et al., 2017). Dan komt er kennis extractie voor TTS en ASR. De TTS-extractie bestaat uit drie stappen. Eerst wordt er voor elke niet gepaarde tekst een corresponderende spraak gemaakt met behulp van het TTS-model en het maken van een

pseudocorpus. Dan wordt het pseudocorpus gefilterd en wordt tenslotte het gefilterde corpus gebruikt om een nieuw TTS-model te trainen. De ASR-extractie bestaat ook uit drie stappen. Eerst moet er een corresponderende tekst bij een ongepaard stukje spraak worden gemaakt door middel van het ASR-model. Dit paar wordt ook in een pseudocorpus gezet, zoals bij de TTS-extractie. Dan wordt er voor elk ongepaard stuk tekst een stuk spraak gemaakt door middel van het TTS-model en dit wordt in een ander pseudocorpus gezet. Tenslotte worden deze twee corpussen gecombineerd en wordt er een nieuw ASR-model getraind. Om de ASR extra te verbeteren wordt SpecAugment gebruikt. LRSpeech is beter dan het model dat in het eerdere onderzoek is gebruikt. De nauwkeurigheid van de ASR is nog niet goed genoeg en willen ze gaan verbeteren in de toekomst. De WER van LRSpeech is 28,82%. Voor een model dat met weinig data werkt, is deze WER goed. Voor een goed werkende ASR is dit echter nog te hoog.

Voor de zero resource methode wordt veel onderzoek gedaan en worden veel pogingen gedaan voor een ASR met een lage WER. Voorlopig zijn deze nog niet zo goed als ASR waar getraind wordt met veel meer taalkundige informatie (Jansen et al., 2013). Het *end-to-end* systeem moet verbeterd worden. Als het model versterkt wordt, kan de correlatie en complementariteit tussen geluid en taal door het model beter worden herkend. Verder zou er kunnen worden gekeken om afbeeldingen en plaatjes te gebruiken om een ASR beter te leren herkennen volgens Yu et al. (2020), omdat dit makkelijker te verkrijgen is en makkelijker is om te labelen. Het systeem van Xu et al. (2020) is interessant om ook in de gaten te houden of deze kan worden verbeterd naar een hogere nauwkeurigheid voor de ASR. Dit zou namelijk betekenen dat er een ASR en TTS zou kunnen worden gemaakt met weinig data. De slechte nauwkeurigheid is een groot nadeel voor het maken van een ASR voor IDS op deze manier.

6. Aanbeveling

Tabel 1

In de volgende tabel worden de voor- (+) en nadelen (-) van de drie benaderingen weergegeven.

Benadering	Voldoende data excl. transcriptie	Voldoende data incl. transcriptie	WER ≤ 20 %	Gebruik van Kaldi mogelijk	Gebruik van G2P regels
Benadering 1	-	+	+	+	+
Benadering 2	+	+	-	+	+
Benadering 3	+	+	-	-	-

Om alle methodes op een rij te hebben, worden de grootste voor- en nadelen nog even noemen van elke benadering en zijn ze samengevat in Tabel 1. Voor de eerste benadering om een nieuwe ASR te maken voor IDS werd er veel gebruik gemaakt van Kaldi. Het voordeel van het gebruik van Kaldi is dat het een stappenplan heeft en makkelijker in gebruik is dan andere toolkits. Ook is het makkelijker dan een hele ASR zelf maken. Het nadeel van alleen Kaldi is dat er meer data nodig zijn dan dat er op het moment is voor een ASR voor IDS om een lage WER te bereiken. Alleen Kaldi en de data van IDS die er nu zijn, zijn niet genoeg. Als Kaldi wordt gecombineerd met SpecAugment, wordt de data vergroten en is de hoeveelheid data groter die in Kaldi kan worden gestopt.

In de tweede benadering is er gekeken naar ASR voor dialecten. De onderzoeken die zijn bekeken, hebben allemaal een nieuwe ASR gemaakt voor het dialect. Alleen Hirayama et al. (2015) weken af en gebruikten meerdere bestaande ASR om te combineren voor een nieuwe ASR. Zo kon deze ASR verschillende dialecten herkennen. Voor IDS zijn deze bestaande ASR er nog niet, dus valt deze methode af. In deze benadering viel de ASR voor het Tunesische dialect op. Deze heeft de meeste voordelen voor het maken van een ASR voor IDS. Hier wordt Kaldi gebruikt met G2P regels en een lexicon van uitzonderingen. De data aan spraak die ze gebruiken is erg beperkt, terwijl de WER ook laag blijft. Het voordeel hiervan is dat er al genoeg spraak is van IDS om deze methode te gebruiken, dus er hoeft geen spraak meer te worden getranscribeerd met de hand. Er moeten nog wel G2P regels worden gemaakt en een lexicon met uitzonderingen. De methode van het one-pass kader voor het Mandarijn en Thais heeft de focus vooral op het herkennen van twee verschillende talen in een zin. Dit is niet nodig voor het IDS.

De laatste benadering kijkt naar alle methodes die zero-resource gebruiken. Hoewel deze methode veelbelovend is, is het nog niet goed genoeg om er nu een ASR mee te maken voor IDS. Ze worden wel steeds beter. De methodes die gebruikt zijn voor zero resource, zijn interessant om in de gaten te houden voor als ze verder ontwikkelen. Het onderzoek van Ren et al. (2019) had een lage WER bij fonemen. Deze zouden ook beter kunnen worden door meer paren te geven als de ASR aan het trainen is. Het vervolgonderzoek heeft LRSpeech gemaakt (Yu et al., 2020). Dit heeft nog minder data nodig om te leren, maar de nauwkeurigheid is nog niet goed voor de ASR. Deze methode is wel interessant om in de gaten te houden als deze verbetert. Dit zou een ASR met TTS maken die maar 20 minuten spraak nodig heeft.

Na afloop van dit onderzoek wordt er aanbevolen om een nieuwe ASR te maken voor IDS aan de hand van Kaldi en MFCC. In Tabel 1 is te zien dat er geen één benadering alleen maar voordelen heeft. De aanbeveling van dit onderzoek is daarom een combinatie van de eerste twee benaderingen. Om de WER zo laag mogelijk te houden, kan er worden gekozen om de data van spraak met transcriptie te vergroten door meer spraak met de hand te transcriberen. Dit zal een lagere WER opleveren. Om minder tijd kwijt te zijn, is een andere aanbeveling om Kaldi te gebruiken en de hoeveelheid data te vergroten met SpecAugment. Om de taalkennis van de ASR te vergroten kunnen er nog G2P regels met behulp van Sequitur worden toegevoegd. Hierop kan de ASR worden getraind en kan er een lexicon met uitzonderingen worden toegevoegd. Aangenomen aan de hand van eerdere onderzoeken zou dit voor een WER van maximaal 22,6 % moeten zorgen. De ASR van het Tunesische dialect gebruikte de combinatie van Kaldi met G2P regels en een lexicon en had een WER van 22,6%. Als er meer data worden toegevoegd, zou dit lager moeten worden zoals de ASR voor het Duits.

7. Discussie

In dit onderzoek zijn verschillende benaderingen voor het maken van een ASR voor IDS bekeken. Vervolgens is er een aanbeveling gedaan door een combinatie te maken van eerdere methodes die gebruikt zijn voor een ASR. Deze aanbeveling is gebaseerd op de aanname dat de ASR beter wordt door Kaldi met een middelmatige hoeveelheid aan data, G2P regels en een lexicon met uitzonderingen dan een ASR op deze manier met een kleinere dataset. Deze methode is nog niet getest. Verder zijn in dit onderzoek niet alle mogelijkheden voor soorten ASR bekeken. Er kan ook nog worden gekeken naar het gebruik van een emotie herkenning (*speech emotion recognition,* SER). IDS zit vol emotie, dus deze methode zou ook kunnen helpen voor het maken van een ASR voor IDS. In onderzoek van Fayek et al. (2016) is er gekeken naar hoe een SER

en ASR elkaar kunnen verbeteren. De conclusie was dat de systemen elkaar kunnen verbeteren, maar dat er nog wel onderzoek naar nodig is.

8. Referenties

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The Journal of the Acoustical Society of America*, *141*(5), 3070–3078. https://doi.org/10.1121/1.4982246

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Biadsy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition* (Theses). https://doi.org/10.7916/D8M61S68

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434–451. https://doi.org/10.1016/j.specom.2008.01.002

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. https://doi.org/10.1016/s0010-0277(01)00122-6

Brownlee, J. (2020, 19 februari). *A Gentle Introduction to Long Short-Term Memory Networks by the Experts*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

Chen, C. M. (2004). Two Types of Code-Switching in Taiwan. *15th Sociolinguistics Symposium*.

Cooper, R. P., & Aslin, R. N. (1994). Developmental Differences in Infant Attention to the Spectral Properties of Infant-Directed Speech. *Child Development, 65*(6), 1663–1677. https://doi.org/10.2307/1131286

Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Processing*, *7*(1), 25–46. https://doi.org/10.1049/iet-spr.2012.0151

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. https://doi.org/10.1121/1.1613311

Doetsch, P., Zeyer, A., Voigtlaender, P., Kulikov, I., Schluter, R., & Ney, H. (2017). Returnn: The RWTH extensible training framework for universal recurrent neural networks. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Published. https://doi.org/10.1109/icassp.2017.7953177

Dunbar, E., Cao, X.-N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., & Dupoux, E. (2017). The Zero Resource Speech Challenge 2017. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 323–330.

Fayek, H. M., Lech, M., & Cavedon, L. (2016). On the Correlation and Transferability of Features Between Automatic Speech Recognition and Speech Emotion Recognition. *Interspeech 2016*. Published. https://doi.org/10.21437/interspeech.2016-868

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. https://doi.org/10.1016/s0163-6383(85)80005-9

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

Ghai, S., & Sinha, R. (2015). Pitch adaptive MFCC features for improving children's mismatched ASR. *International Journal of Speech Technology*, *18*(3), 489–503. https://doi.org/10.1007/s10772-015-9291-7

Guglani, J., & Mishra, A. N. (2018). Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. *International Journal of Speech Technology*, *21*(2), 211–216. https://doi.org/10.1007/s10772-018-9497-6

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*(4), 1738–1752. https://doi.org/10.1121/1.399423

Hirayama, N., Yoshino, K., Itoyama, K., Mori, S., & Okuno, H. G. (2015). Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(2), 373–382. https://doi.org/10.1109/taslp.2014.2387414

Hori, T., Watanabe, S., Zhang, Y., & Chan, W. (2017). Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM. *Proc. Interspeech*, 949–953.

HTK Speech Recognition Toolkit. (2016). HTK Speech Recognition Toolkit. Geraadpleegd op 18 juni 2021 van https://htk.eng.cam.ac.uk/

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., Feldman, N., Hermansky, H., Metze, F., Rose, R., Seltzer, M., Clark, P., McGraw, I., Varadarajan, B., Bennett, E., Borschinger, B., Chiu, J., Dunbar, E., Fourtassi, A., . . . Thomas, S. (2013). A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/icassp.2013.6639245

Ketkar, N. (2017). *Deep Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-2766-4

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. *Proc. ICASSP*, 181–184.

Kozierski, P., Sadalla, T., Drgas, S., Dąbrowski, A., & Horla, D. (2016). Kaldi Toolkit in Polish Whispery Speech Recognition. *Przegląd Elektrotechniczny*, *1*(11), 303–306. https://doi.org/10.15199/48.2016.11.70

Lee, A., Kawahara, T., & Shikano, K. (2001). Julius - An Open Source Real-Time Large Vocabulary Recognition Engine. *EUROSPEECH2001: the 7th European Conference on Speech Communication and Technology*, 1691–1694.

Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1), 35–45. https://doi.org/10.1109/29.45616

Lyu, D., Lyu, R., Chiang, Y., & Hsu, C. (2006). Speech Recognition on Code-Switching Among the Chinese Dialects. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. https://doi.org/10.1109/icassp.2006.1660218

Macwhinney, B. (2000). *The Childes Project: Tools for Analyzing Talk* (Third Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., & Belguith, L. (2017). Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation*, *52*(1), 249–267. https://doi.org/10.1007/s10579-017-9402-y

Milde, B., & Koehn, A. (2018). Open Source Automatic Speech Recognition for German. *Speech Communication; 13th ITG-Symposium*, 1–5. https://arxiv.org/pdf/1807.10311.pdf

Open Source Speech Recognition Toolkit. (2019, 23 oktober). CMUSphinx. Geraadpleegd op 18 juni 2021 van https://cmusphinx.github.io/

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*. https://doi.org/10.21437/interspeech.2019-2680

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. *Proceedings of The future of gradientbased machine learning software and techniques (Autodiff) in the twenty-ninth annual conference on neural information processing systems (NIPS)*.

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers Consistently Alter Their Unique Vocal Fingerprints When Communicating with Infants. *Current Biology*, *27*(20), 3162–3167.e3. https://doi.org/10.1016/j.cub.2017.08.074

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. *Proc. ASRU*.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., & Biermann, C. (2015). Open Source German Distant Speech Recognition: Corpus and Acoustic Model. *Proc. Text, Speech, and Dialogue (TSD)*, 480–488.

Reiko, M., Yosuke, I., & Ken'ya, N. (2006). Input for Learning Japanese : RIKEN Japanese Mother-Infant Conversation Corpus. *IEICE technical report*, *106*, 11–15.

Ren, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. (2019). Almost Unsupervised Text to Speech and Automatic Speech Recognition. *International Conference on Machine Learning*, 5410–5419.

Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., & Ney, H. (2009). The RWTH Aachen University Open Source Speech Recognition System. *Interspeech*, 2111–2114.

Stolcke, A. (2002). SRILM-an Extensible Language Modeling Toolkit,. *In Proc. Intl. Conf. Spoken Language Processing (INTERSPEECH)*.

Tokui, S., Oono, K., Hido, S., & Clayton, J. (2015). Chainer: a Next-Generation Open Source Framework for Deep Learning. *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, *5*.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *PNAS*, *104*(33), 13273–13278. https://doi.org/10.1073/pnas.0705369104

van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2020). Automatic Recognition of Target Words in Infant-Directed Speech. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 522. https://doi.org/10.1145/3395035.3425184

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Neural Information Processing Systems (NIPS)*. Published.

Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The Zero Resource Speech Challenge 2015: Proposed Approaches and Results. *Procedia Computer Science*, *81*, 67–72. https://doi.org/10.1016/j.procs.2016.04.031

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning - ICML '08*. https://doi.org/10.1145/1390156.1390294

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., & Ochiai, T. (2018). ESPnet: End-to-End Speech Processing Toolkit. *Proceedings of 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2207–2211.

Wiesner, M., Liu, C., Ondel, L., Harman, C., Manohar, V., Trmal, J., Huang, Z., Dehak, N., & Khundanpur, S. (2018). Automatic Speech Recognition and Topic Identification for Almost-Zero-Resource Languages. *Proc. Interspeech*. Published.

Wikipedia contributors. (2021, 13 mei). *Wikipedia:WikiProject Spoken Wikipedia - Wikipedia*. Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia

Xu, J., Tan, X., Ren, Y., Qin, T., Li, J., Zhao, S., & Liu, T. Y. (2020). LRSpeech. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3394486.3403331

Young, S. (1997). *The HTK Book*. Cambridge University Press.

Yu, C., Kang, M., Chen, Y., Wu, J., & Zhao, X. (2020). Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview. *IEEE Access*, *8*, 163829–163843. https://doi.org/10.1109/access.2020.3020421

Zhang, Q., & Liu, X. (2020). Improve Mandarin Speech Recognition Using Simple Recurrent Units and SpecAugment. *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. Published. https://doi.org/10.1109/auteee50969.2020.9315706

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589. https://doi.org/10.1007/bf02943243