HO
GENT

Faculteit Bedrijf en Organisatie

Nursery Tone Monitor: softwarematige detectie van elderspeak

Glenn Beeckman

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Koen Mertens
Co-promotor:
Jorrit Campens

Instelling: HOGENT

Academiejaar: 2020-2021

Tweede examenperiode

Faculteit Bedrijf en Organisatie

Nursery Tone Monitor: softwarematige detectie van elderspeak

Glenn Beeckman

Scriptie voorgedragen tot het bekomen van de graad van
professionele bachelor in de toegepaste informatica

Promotor:
Koen Mertens
Co-promotor:
Jorrit Campens

Instelling: HOGENT

Academiejaar: 2020-2021

Tweede examenperiode

Woord vooraf

Het schrijven van deze bachelorproef bleek één van de grotere uitdagingen van mijn opleiding 'Toegepaste Informatica' aan de hogeschool Gent. Ik koos dit onderwerp omdat dit aanleunde bij mijn interesses en dit een mooie aansluiting was aan mijn vorige opleiding: 'Technoloog Medische Beeldvorming'.

Het combineren van de zorgsector met mijn kennis van IT leek mij uitermate interessant. Het gebruik van elderspeak is een gegeven waar ik vanuit eigen ervaring vaak mee in contact gekomen ben. Dit zijnde door het te horen bij anderen, maar eveneens door mezelf erop te betrappen. Een tool die feedback kan geven op het gebruik hiervan, leek mij handig en het zou dan ook een mooie kans zijn als ik hier een waardige bijdrage aan kan leveren.

Tijdens het maken van deze bachelorproef werd snel duidelijk dat, omwille van covid-19, het zeer moeilijk zou worden om audiosamples te verzamelen. Speciaal om deze reden zou ik mijn vriendengroep willen bedanken voor het doneren van audio-opnames aan dit onderzoek. Zonder deze opnames zou deze bachelorproef nooit tot een volwaardig einde gebracht zijn.

Verdere dank wil schenken aan mijn promotor, Koen Mertens, voor het steunen van deze bachelorproef en het geven van waardige feedback. Bijkomend hielp de heer Mertens met het bedenken van oplossingen die een bijdrage hadden tot het schrijven van deze bachelorproef. Ook zijn bezorgdheid rond mijn mentale gezondheid werd geapprecieerd.

Ten slotte zou ik ook mijn copromotor, Jorrit Campens, willen bedanken voor zijn steun en enthousiasme. Zonder hem zou dit onderwerp er mogelijk niet geweest zijn.

Glenn Beeckman

Samenvatting

In deze bachelorproef is nagegaan of het mogelijk is elderspeak softwarematig te detecteren aan de hand van de toonhoogte en eventueel door middel van machine learning.

Tijdens dit onderzoek werden audio-opnames verzameld die geanalyseerd werden op basis van de toonhoogte. Er werd een python applicatie gemaakt die de optie geeft twee opnames te maken en deze opnames nadien te analyseren naar de toonhoogte. De applicatie maakt gebruik van de 'Parselmouth' bibliotheek voor python, die op zijn beurt gebruik maakt van de 'Praat' source code. De applicatie in python blijkt snel en nauwkeurig de toonhoogte te bepalen en lijkt praktisch in gebruik. Wanneer deze applicatie uitgebreid kan worden met bijkomende functies, zoals het detecteren van herhaalde woorden/zinnen en een detectie van verkleinwoorden, zou dit een volwaardige manier kunnen zijn om elderspeak softwarematig te detecteren.

Deze bachelorproef heeft ook geprobeerd elderspeak te detecteren aan de hand van een convolutioneel neuraal netwerk (CNN). Het CNN gebruikte een te kleine dataset van slechts 20 personen (40 audiosamples in totaal) waardoor het onmogelijk was een nauwkeurige classificatie te maken. Zo had het model een accuraatheid die schommelt tussen de 45 en 60 procent. Aangezien er 'slechts' twee klassen waren is dit niet voldoende omdat gokken statistisch gezien ook een nauwkeurigheid van 50 procent zou halen. Het toevoegen van geaugmenteerde data had in dit onderzoek geen effect op de accuraatheid van het model. Over het aspect van het CNN zou verder onderzoek gevoerd kunnen worden om na te gaan of dit eventueel een praktische manier is om elderspeak te gaan detecteren. Hiervoor zou echter een grotere en meer gevarieerde dataset nodig zijn.

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

1. Inleiding

1.1 Context

Wereldwijd neemt het aandeel ouderen in de bevolking spectaculair toe. In 1990 was 9,2 procent van de wereldbevolking zestig jaar of ouder, terwijl verwacht wordt dat tegen 2050 maar liefst 21,1 procent van de wereldbevolking minstens zestig jaar zal zijn. Ook in Vlaanderen wordt deze trend gevolgd. In 2000 was het aandeel van zestigplussers 16,7 procent. In 2020 was dit aandeel met 3,8 procent gestegen tot 20,5 procent. Volgens de huidige statistiek lijkt deze trend zich naar de toekomst toe ook door te zetten (Bynens, 2020). In deze vergrijzende context, komen zorgverleners meer en meer in contact met deze oudere zorgvragers. Dit geldt zowel voor zorgverleners die werkzaam zijn in woonzorgcentra, maar evengoed voor zorgverleners die werken binnen een thuiscontext.

1.2 Nood

Aangezien een goed sociaal contact een positieve invloed heeft op de mentale gezondheid van een individu, is het belangrijk dat de behoefte aan sociaal contact bij de ouderen wordt vervuld. Toch wordt deze manier van communiceren vaak als ontoereikend beschouwd door zowel de zorgvrager als zorgverlener (Balsis & Carpenter, 2006). Hoewel het gebruik van elderspeak door een beperkte groep ouderen als positief kan worden ervaren, wordt het door de meerderheid eerder als negatief beschouwd. Zo wordt het vaak ervaren als onrespectvol en vinden de ouderen de instructies die gegeven worden in elderspeak verwarrend of onduidelijk. Verder wordt de communicatie door deze ouderen ervaren als moeizaam (Lowery, 2013). Om het mentale welzijn van de oudere zo hoog mogelijk te houden, moet het gebruik van elderspeak vermeden worden.

1.3 Probleemstelling

Om zorgverleners meer bewust te maken van het al dan niet gebruiken van elderspeak, is het belangrijk een goede tool beschikbaar te stellen die feedback kan geven op het gebruik hiervan. Wanneer deze tool gebruikt kan worden tijdens opleiding van een zorgverlener, kan de zorgverlener snel bewust gemaakt worden van het al dan niet gebruiken van elderspeak. Zo kan deze manier van spreken vermeden worden in de carrière van de zorgverlener. Nadien zou deze tool eventueel uitgebreid kunnen worden naar een bredere context dan de zorgsector.

1.4 Onderzoeksvraag

Deze bachelorproef zal zich focussen op de vraag of het mogelijk is elderspeak softwarematig te detecteren aan de hand van de toonhoogte. Er zal in de eerste plaats nagegaan worden of het mogelijk is elderspeak te detecteren aan de gemiddelde toonhoogte in een gesproken audiosample. Vervolgens zal nagegaan worden of een correcte classificatie kan gemaakt worden door het trainen van een neuraal netwerk.

1.5 Onderzoeksdoelstelling

Deze bachelorproef hoopt aan te tonen dat de toonhoogte een goede parameter is waarop elderspeak softwarematig gedetecteerd kan worden. In deze bachelorproef zal een praktisch voorbeeld uitgewerkt worden om na te gaan of de gemiddelde toonhoogte berekend kan worden binnen een reële context.

1.6 Opzet van deze bachelorproef

In Hoofdstuk 2 van deze bachelorproef wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie. Er zal informatie gegeven worden over: elderspeak, de toonhoogte, het Praat programma, informatie over de gebruikte python-bibliotheek en een korte inleiding over neurale netwerken alsook hoe deze gebruikt kunnen worden in het classificeren van audio.

In Hoofdstuk 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen. Hier is onder andere informatie te vinden over de audiosamples alsook op welke manier er te werk gegaan is in het verwerken van deze data.

Hoofdstuk 4 behandelt kort de resultaten met betrekking tot de toonhoogte en het neurale netwerk.

Hoofdstuk 5, tenslotte, formuleert een conclusie en biedt het een antwoord op de onderzoeksvragen. In dit hoofdstuk wordt eveneens een aanzet gegeven over de toekomst van dit project en mogelijk vervolgonderzoek dat uitgevoerd zou kunnen worden.

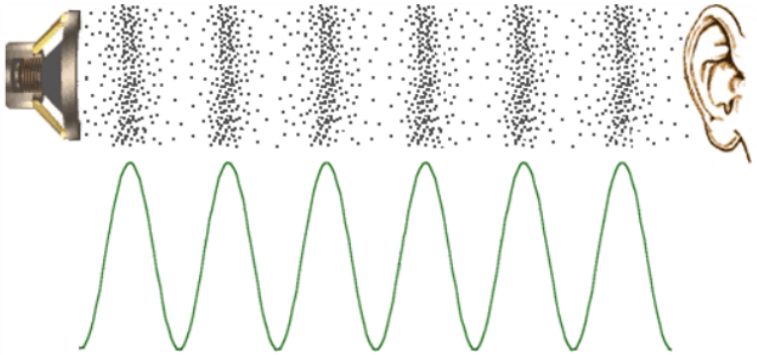
2. Stand van zaken

Inleiding

Aangezien deze bachelorproef zal trachten elderspeak softwarematig te detecteren, is het belangrijk kort stil te staan bij wat elderspeak precies inhoudt. Aangezien het iets is waar de meeste mensen (al dan niet bewust) mee te maken hebben en er mogelijk hun eigen invulling aan geven, is het nodig dit te definiëren. Dit hoofdstuk zal verder ingaan op de toonhoogte en de detectiemogelijkheden hiervan. Ten slotte wordt kort vermeld hoe neurale netwerken een rol kunnen spelen in het detecteren van elderspeak.

2.1 Elderspeak

Elderspeak is een vorm van communicatie waarbij de spreker gebruikt maakt van een zeer simpel taalgebruik. Dit taalgebruik kenmerkt zich vooral door een simpelere grammatica en woordenschat. Ook zal de spreker meer variëren in zijn/haar toonhoogte en zal de spreker met een hoger volume communiceren. Elderspeak is een communicatiestijl die voornamelijk wordt gebruikt wanneer een jongere volwassene spreekt tegenover een oudere volwassene (Kemper, Finter-Urczyk e.a., 1998). De reden waarom deze stijl wordt gebruikt, ligt bij het stereotiep waarbij de oudere wordt bestempeld als ‘minder begaafd‘ en minder snel van begrip. Dit stereotiep wordt versterkt wanneer deze oudere extra zorgen nodig heeft en/of wanneer er sprake is van neurologische aandoeningen zoals dementie of parkinson (Balsis & Carpenter, 2006).



Figuur 2.1: Voorbeeld van hoe een geluidsgolf zich voortbeweegt door een medium (Hart, 2011).

2.2 Toonhoogte

Aangezien de toonhoogte een zeer belangrijke eigenschap is van elderspeak, is het nodig te duiden wat hier precies mee bedoeld wordt. Een geluidsgolf zal zich (zoals de naam reeds aangeeft) voortbewegen met een golfbeweging. Deze golf beweegt zich voort door middel van een medium (in het dagelijkse leven zal dit medium vaak bestaan uit lucht). Dit medium zal door een geluidsbron mee trillen en deze trillingen kunnen door een bepaalde ontvanger geïnterpreteerd worden als geluid. Figuur 2.1 toont aan hoe een golf zich voortbeweegt door een medium. De golf zal zich door dit medium voortbewegen met een bepaalde frequentie. De frequentie van een golf duidt aan hoe vaak de deeltjes van het medium mee vibreren wanneer een golf door dit medium passeert (The Physics Classroom, g.d.). De frequentie van een golf wordt uitgedrukt in Hertz (Hz), waarbij:

$$1Hertz = \frac{1Vibratie}{Seconde}$$

Wanneer geluiden op een organische manier geproduceerd worden (bv. door menselijke stembanden), zal deze geluidsgolf niet altijd even harmonisch zijn. Er zullen lichte fluctuaties liggen in de geluidsgolf wat ervoor zorgt dat er verschillende frequenties aanwezig zijn binnen eenzelfde geluidsgolf. Om deze fluctuaties op te vangen wordt de toonhoogte ook vaak uitgedrukt als ‘de fundamentele frequentie‘ of ‘de grondfrequentie‘ (F0). Deze F0 is de laagste frequentie die in een signaal voorkomt en wordt gebruikt om de toonhoogte van dit geluid te bepalen en uit te drukken in Hertz (Francart & Eneman, 2008). Het is deze F0 die in deze bachelorproef gedetecteerd zal worden.

2.3 F0 detectie

2.3.1 Overzicht

Informatie vergaren over de F0 kan zeer handig zijn in het analyseren van spraakgerelateerde audio. Zo wordt het gebruikt bij de analyse van linguïstische en pragmatische functies. Bij real time detectie hiervan kan het informatie geven over de intenties van

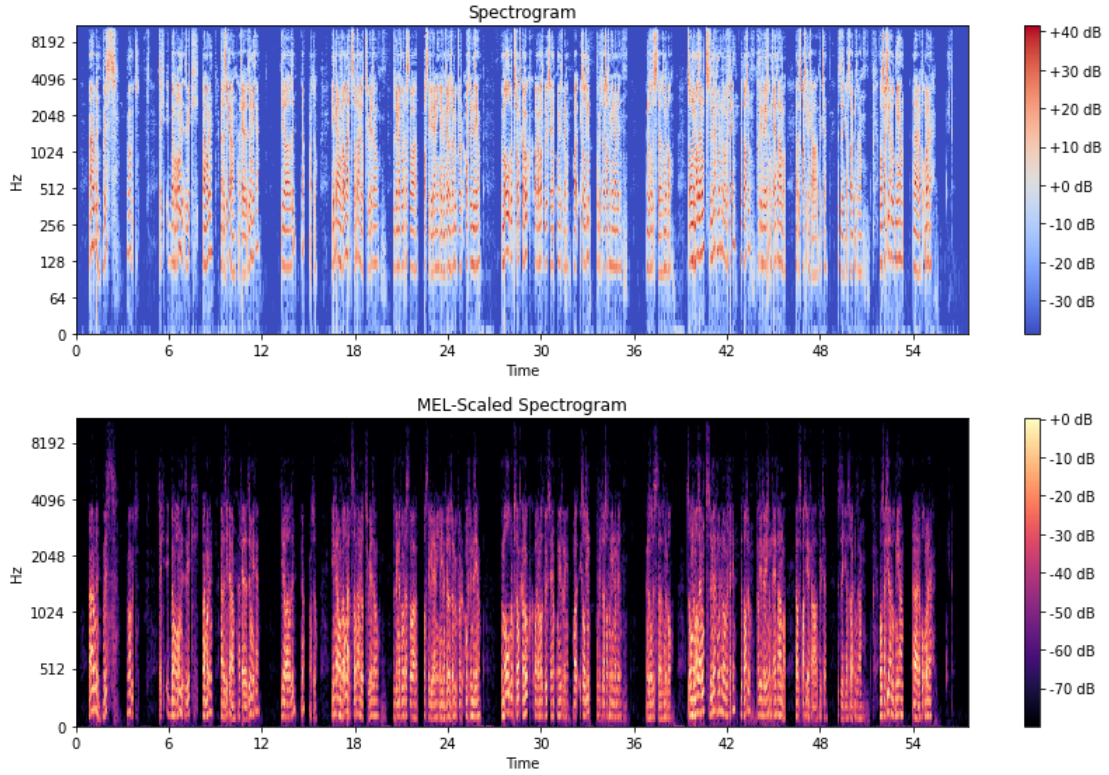
de spreker. Voor de detectie van deze F0 bestaan verschillende methoden. Voorbeelden hiervan zijn: een auto-correlatie functie (ACF), het robuuste algoritme voor pitch tracking (RAPT), het YIN-algorimte en de tijd-domein excitatie extractie gebaseerd op een minimum verstoring operator (TEMPO). Andere maken gebruik van de zaagtandgolf geïnspireerde pitch estimator (SWIPE). Er kan ook gebruik gemaakt worden van het nearly free F0 estimation algoritme (NDF). De meest gebruikte algoritmen zijn: Praat, RAPT, STRAIGHT, YIN en SWIPE (Strömbergsson, 2016). Wanneer deze algoritmen met elkaar vergeleken werden, bleek dat het Praat algoritme één van de laagste F0 Frame Errors (FFE) maakte op gesproken data (Jouvet & Laprie, 2017). Deze FFE duidt op het aantal frames waarin fouten rond de toonhoogte gemaakt werden, en kan aanzien worden als een meeteenheid om de performantie van een pitch-tracker te controleren. Dit algoritme had ook goede resultaten wanneer de datasets een zekere ruis hadden (Jouvet & Laprie, 2017). Dit is goed aangezien de 'Nursery Tone Monitor' gebruikt zal worden in reële situaties waar een zekere ruis mogelijk aanwezig zal zijn.

2.3.2 Spectrogram

Een spectrogram is een manier waarop audio weergegeven kan worden. Het toont de signaalsterkte en de verschillende frequenties die aanwezig zijn in een bepaalde golf. Spectrogrammen kunnen tweedimensionaal zijn waarin de frequenties van een golf uitgezet worden in functie van de tijd. Soms bestaat een spectrogram uit een derde dimensie die de energie of de luidheid van een golf weergeeft. Deze derde dimensie wordt weergegeven aan de hand van kleuren waar een rode kleur overeenkomt met een hoge amplitude en een blauwe kleur overeenkomt met een lage amplitude (Kamp, 2020).

2.3.3 MEL frequentie spectrogram

Een gewoon spectrogram is voor dit onderzoek echter niet voldoende. De reden hiervoor ligt in de manier hoe mensen geluid (en meer specifiek de frequentie) interpreteren. Het menselijk gehoor neemt frequenties niet-lineair waar. Een gewoon spectrogram geeft geluid echter wel op deze manier weer. De spectrogrammen zullen dus nog omgezet moeten worden naar een mel-spectrogram. Mel-frequentie spectrogrammen zijn spectrogrammen waarbij de frequentieschaal aangepast is naar deze van het menselijk gehoor en zijn gebaseerd op de mel-schaal ('melody-schaal'). De mel-frequentie spectrogrammen zijn zodanig opgesteld dat frequenties gepositioneerd zijn, als hoe mensen deze zouden ervaren (Ramaseshan, 2013). Mensen zijn namelijk beter in het onderscheiden van lage tonen dan ze zijn in het onderscheiden van hoge tonen. Figuur 2.2 toont het verschil aan tussen een normaal spectrogram (bovenaan) tegenover een mel-frequentie spectrogram (onderaan). Het belangrijkste verschil tussen de twee is zoals eerder vermeld, de schaal waarop de frequenties weergegeven worden.

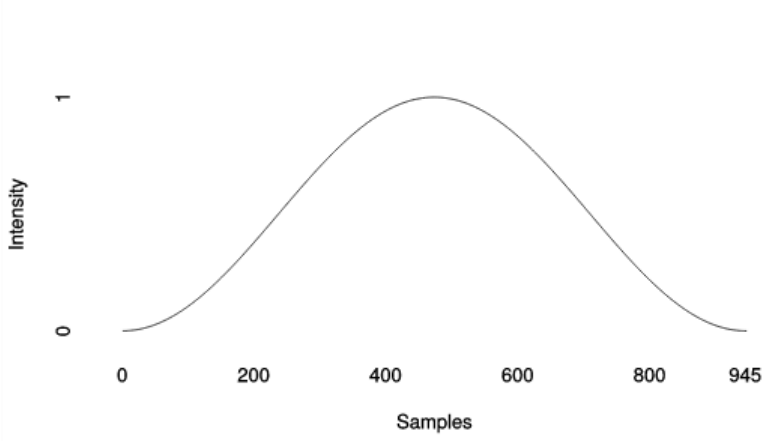


Figuur 2.2: Normaal spectrogram (bovenaan) vergeleken met een mel-frequentie spectrogram (onderaan).

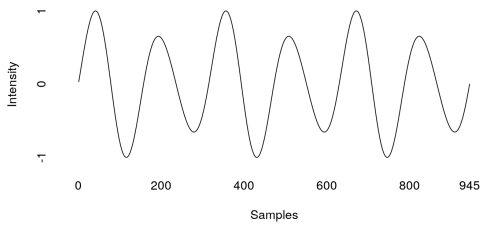
2.4 PRAAT

Praat is een freeware programma dat ontwikkeld is door ‘Paul Boersma‘ en ‘David Weenink‘ van de Universiteit van Amsterdam. Het programma kan gebruikt worden om gesproken akoestische signalen te analyseren (van Lieshout, 2003). De moeilijkheid in het analyseren van de toonhoogte in gesproken taal, zit in het feit dat gesproken taal niet bestaat uit één continue toon. Ieder deel ervan bestaat uit een combinatie van verschillende tonen wat het moeilijk maakt een goede analyse ervan te maken. Praat overkomt dit door te beweren dat spraak stabiel genoeg is wanneer er wordt gekeken naar fragmenten die klein genoeg zijn. Deze kleine fragmenten worden ook wel ‘het venster van de analyse‘ genoemd. Om de analyse beter te maken, worden deze vensters gefilterd aan de hand van een hanningfunctie (Boersma, 1993). Door de vensters te vermenigvuldigen met de hanningfunctie (figuur 2.3), worden de pieken aan de uiteinden van deze vensters weg gefilterd.

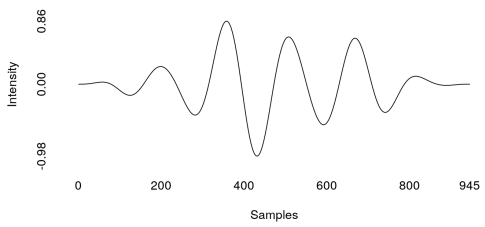
Om toonhoogte te detecteren, wordt er bij Praat gebruikt gemaakt van autocorrelatie. Deze autocorrelatie kan er echter voor zorgen dat in sommige gevallen de verkeerde piek (of toonhoogte in ons geval) binnen een golf wordt gedetecteerd. Om dit probleem tegen te gaan, zal het algoritme het gefilterde signaal delen door de genormaliseerde autocorrelatie van de vensterfunctie. Het resultaat van deze bewerking geeft een goede weergave van de geschatte toonhoogte en is zeer performant in het achterhalen van de toonhoogte in gesproken audio (Boersma, 1993).



Figuur 2.3: weergave van een Hanningfunctie. (Atria, 2015)

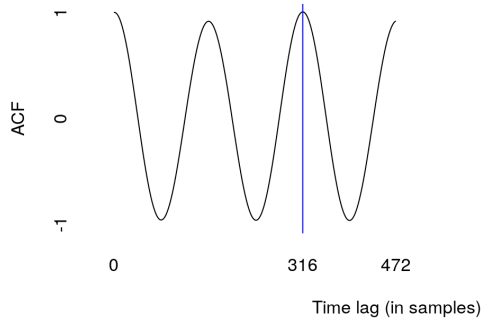


(a) Drie cyclussen van een complexe sinusgolf met een piek amplitude van 1, beginnend en eindigend bij 0. De horizontale as is 945 samples lang.



(b) Een Hanning-gefilterde complexe sinus golf met een amplitude dicht bij 0 aan de uiteinden en 1 in het midden.

Figuur 2.4: Sinusfunctie voor en na filtering d.m.v. Hanning functie. (Atria, 2015)



Figuur 2.5: Drie cyclussen van een sinusgolf met de helft van de lengte als de vorige. De eerste heeft een amplitude van 1, de tweede is licht lager dan de eerste. De derde piek, die licht hoger is als de tweede, is gemarkeerd met een blauwe lijn op sample 316. (Atria, 2015)

Vervolgens kan uit deze samples de toonhoogte berekend worden aan de hand van volgende formule:

$$f_0 = \frac{1}{lag_{max}/f_s}$$

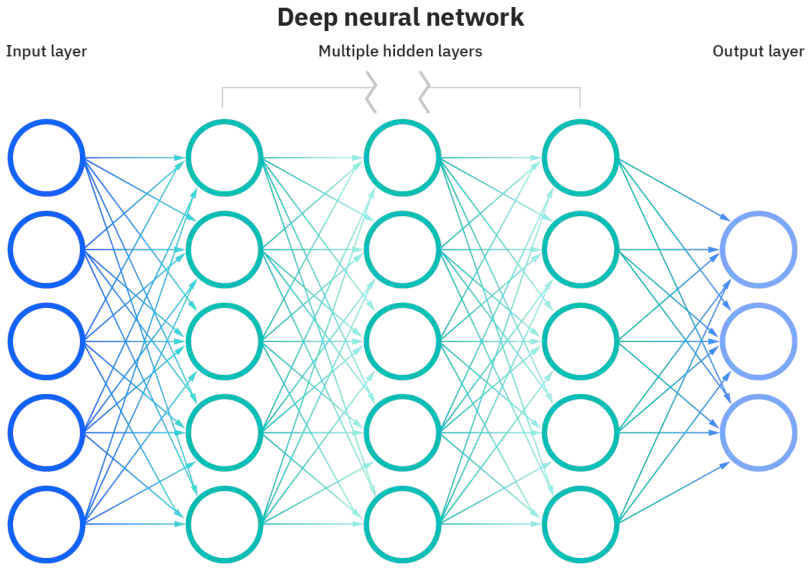
2.5 Parselmouth

Parselmouth is een relatief nieuwe bibliotheek voor python die het mogelijk maakt Praat functionaliteiten te gebruiken in Python. De bibliotheek is nog in ontwikkeling en er worden op regelmatige basis nieuwe functionaliteiten van Praat toegevoegd aan deze bibliotheek. Parselmouth onderscheidt zich van gelijkaardige bibliotheken door rechtstreeks gebruik te maken de Praat open-source code die geschreven werd in C/C++. Andere bibliotheken gebruiken hiervoor vaak de Praat-scripts. Door gebruik te maken van de source-code, is het mogelijk Praat functionaliteiten te gebruiken aan de hand van Python syntax, zonder dat de gebruikers een andere scripting taal moeten aanleren. Dit vergemakkelijkt het gebruik van deze bibliotheek tegenover zijn tegenhangers. Een bijkomend voordeel van rechtstreeks gebruik te maken van de source-code, is dat de verwerking van audiobestanden snel en efficiënt kan gebeuren (Jadoul e.a., 2018).

2.6 Neurale netwerken

Neurale netwerken bestaan uit een serie van algoritmen die trachten onderlinge relaties te herkennen in een bepaalde dataset. Het proces waarop een neuraal netwerk dit doet, is gelijkaardig aan de manier dat een menselijk brein dit probleem zou aanpakken (aan de hand van neuronen). Het voordeel van neurale netwerken is dat deze zich kunnen aanpassen wanneer de input verandert. Ongeacht van de input data, zal een neuraal netwerk steeds proberen een zo goed mogelijk resultaat (of output) te voorspellen (Chen, 2020). Een neuraal netwerk zal dit proberen doen door iedere node binnen zijn lagen een verschillende lineaire regressie formule te laten toepassen. Iedere node binnen een bepaalde laag zal dus bestaan uit een unieke combinatie van de inputgegevens. In het begin zullen dit relatief voor de hand liggende combinaties zijn, maar naargelang het neurale netwerk dieper wordt, zullen deze combinaties abstracter worden en hopelijk leiden tot een zo correct mogelijke inschatting van de output.

Het belang die iedere inputparameter heeft op de output, wordt bepaald aan de hand van gewichten. Zo kan parameter ‘a‘ bijvoorbeeld veel meer meetellen in de formule dan parameter ‘b‘. Wanneer het neurale netwerk bestaat uit meerdere lagen, zal de output van de ene laag dienen als input voor de volgende laag. Dit wordt vaak verwezenlijkt door middel van een activatiefunctie. Wanneer de output van een bepaalde node een voorafbepaalde drempelwaarde overschrijdt, zal de data van deze node doorgegeven worden naar de volgende laag (IBM Cloud Education, 2020). Figuur 2.6 toont een voorbeeld van een neuraal netwerk. Iedere node in deze figuur staat gelijk aan een specifieke combinatie van de inputparameters.



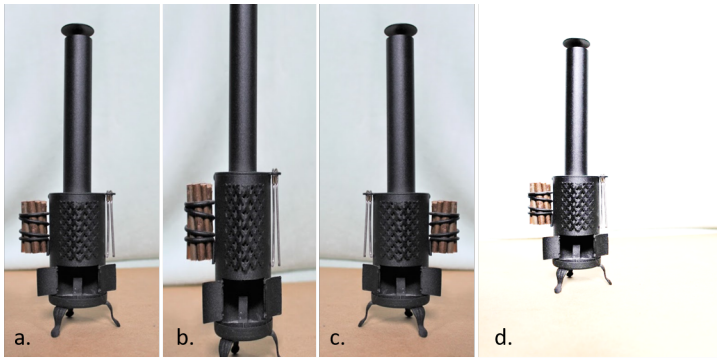
Figuur 2.6: Voorbeeld van een neuraal netwerk met een inputlaag, verscheidene verborgen lagen en een output laag. (IBM Cloud Education, 2020)

2.6.1 Convolutioneel neuraal netwerk

Een convolutioneel neuraal netwerk (CNN) is gelijkaardig aan een normaal neuraal netwerk met het verschil dat deze een convolutionele laag hebben die het mogelijk maakt afbeeldingen zeer goed te classificeren. Het classificeren van afbeeldingen kan worden gedaan door de afbeelding om te zetten naar een matrix van pixelwaarden. Bij grote datasets met zeer grote resoluties, is het computioneel zeer moeilijk dit te verwerken. Dit is waar de convolutionele laag voornamelijk een nut heeft (Saha, 2018). Als eerste zullen er bepaalde kernels of filters over de afbeelding gaan. Deze filters kunnen afbeeldingen verkleinen en verwerken zonder dat belangrijke kenmerken verloren gaan. Deze filters zijn verschillend van elkaar wat het mogelijk maakt dat iedere filter een bepaalde eigenschap van de afbeelding kan detecteren. In het begin zullen er op deze manier voornamelijk randen gevonden worden (voorbeeld: kernel 1 vindt horizontale randen, kernel 2 vindt verticale randen, etc.). Naarmate er meer convolutionele lagen zijn, zullen de gevonden eigenschappen complexer worden. Zo kunnen combinaties van deze randen ervoor zorgen dat er bv. cirkels, huizen, gezichten, etc. gedetecteerd worden, aangezien deze bestaan uit complexe combinaties van deze randen.

2.6.2 Audio classificatie met CNN

Convolutionele neurale netwerken zijn goed in het classificeren van afbeeldingen. Deze bachelorproef zal trachten audiosamples te analyseren. De audiosamples kunnen wel omgezet worden naar mel-frequentie spectrogrammen die kunnen dienen als input voor een CNN. Onderzoek heeft aangetoond dat door het gebruik van een CNN, deze spectrogrammen geanalyseerd kunnen worden op bepaalde patronen en dat audioclassificatie door



Figuur 2.7: Voorbeeld van data augmentatie waarin een originele afbeelding bewerkt werd met verschillende transformaties. a: bijgesneden origineel, b: crop, c: gespiegeld, d: kleurcorrectie

middel van een CNN mogelijk is. Zo worden nauwkeurigheden behaald die variëren van 66.44 tot 90.51 procent (Nanni e.a., 2021).

2.6.3 Data augmentation

De accuraatheid van een deep-learning model is zeer sterk afhankelijk van de grootte van de dataset die als input gebruikt kan worden. Grote datasets bemachtigen is echter niet altijd evident. Een mogelijke oplossing tot het generen van meer inputdata kan door middel van 'data augmentatie'. Door verschillende transformaties uit te voeren op bestaande datasets, kan deze dataset enorm toenemen in aantal. Bij de classificatie van afbeeldingen zouden de afbeelding bijvoorbeeld ingezoomd, gedraaid, gespiegeld, etc. kunnen worden. Op deze manier 'denkt' het model dat de inputafbeeldingen verschillend zijn, ondanks dat deze bestaan uit transformaties van de originele dataset. Figuur 2.7 toont aan hoe 1 afbeelding vermenigvuldigd kan worden door middel van data augmentatie. Wanneer modellen die getraind werden met gewone dataset vergeleken worden met modellen die geaugmenteerde data tot hun beschikking hadden, blijkt dat de modellen met geaugmenteerde data het meestal beter doen dat de andere modellen. Data augmentatie waarin verschillende transformaties gecombineerd worden, blijken ook de hoogste nauwkeurigheid te bereiken. (Shorten & Khoshgoftaar, 2019). Er moet wel opgemerkt worden dat deze transformaties binnen een realistisch venster toegepast moeten worden. Wanneer de input afbeeldingen niet meer overeenkomen met de realiteit, zal het model ook geen correcte classificatie kunnen maken. Het 'garbage in, garbage out' (GIGO) concept is hier sterk van toepassing.

2.6.4 Data augmentation bij audiosamples

Data augmentatie kan ook toegepast worden bij het verwerken van audiosamples. Hier zullen echter audio-specifieke transformaties moeten gebeuren en moet er rekening gehouden worden dat de getransformeerde audio bruikbaar blijft voor het model. Mogelijke transformaties voor audio zijn: 'uitrekken van de tijd, het licht verplaatsen van het

spectrogram, veranderingen in de toonhoogte, Dynamic Range Compression (DCR) en het toevoegen van ruis‘. Onderzoek toont aan dat het gebruik van geaugmenteerde data meestal leidt tot een stijging in nauwkeurigheid van 6 procent (Salamon & Bello, 2017). Er moet wel benadrukt worden dat dit niet altijd het geval is. In het onderzoek was het model beter in staat de individuele klassen te onderscheiden van elkaar, maar nam de potentiële verwarring tussen bepaalde klassen wel toe na het gebruik van geaugmenteerde data. Het is zeer belangrijk per dataset goed na te denken over welke transformaties er toegepast kunnen worden en welke best niet (Salamon & Bello, 2017).

3. Methodologie

3.0.1 Verzamelen van data

Eerst werden er opnames verzameld rond elderspeak. Aan verschillende deelnemers werd gevraagd drie opnames in te spreken met de volgende instructies:

- casus 1 was een gesprek met een fictieve leeftijdgenoot, vriend of collega. Er werd gevraagd te praten over een daguitstap naar de zee.
- In casus 2 werd gevraagd dezelfde situatie uit te leggen aan een 90-jarige bewoner van een woon-zorgcentrum.
- In casus 3 werd opnieuw gevraagd dezelfde situatie uit te leggen, maar dit maal aan een 90-jarige woon-zorgbewoner met kenmerken van neurologische aandoeningen als dementie of parkinson.

De opnames werden opgeslagen als '.wav' bestanden. Oorspronkelijk was het de bedoeling deze audiosamples te vergaren in woon-zorgcentra en in de opleiding verpleegkunde. Door Covid-19 was het niet mogelijk dit te doen binnen woon-zorgcentra. Aangezien de studenten verpleegkunde niet in staat waren te oefenen in hun praktijklokalen, werd digitaal een oproep gedaan in deze opleiding. Vanuit verpleegkunde werden echter geen opnames verkregen waardoor de opnames vergaard moesten worden bij gewone burgers. Aangezien het niet mogelijk was deze opnames in te laten spreken door zorgpersoneel, werden de deelnemers ook gevraagd in casus twee en drie enkele eigenschappen toe te passen van elderspeak. Omdat de deelnemers zich thuis bevonden, werd dit gedaan door een microfoon die ze zelf ter beschikking hadden. Er werden uiteindelijk opnames verzameld van 20 personen. Deze opnames werden gesorteerd per casus.

3.0.2 Verwerken van de data

De data werd nadien verwerkt met de Praat-software om na te gaan of de opnames variëren op basis van toonhoogte. Deze bevindingen dienen als referentie omdat Praat zeer goed is in het analyseren van gesproken data. De opnames werden omgezet naar een spectrogram met als ondergrens 75Hz en als bovengrens 600Hz. Deze grenzen zorgen ervoor dat stiltes of zeer hoge toonhoogten niet meegenomen worden in de analyse. Vervolgens werd de gemiddelde toonhoogte per opname berekend en opgeslagen. Dit werd voor alle 3 de casussen toegepast.

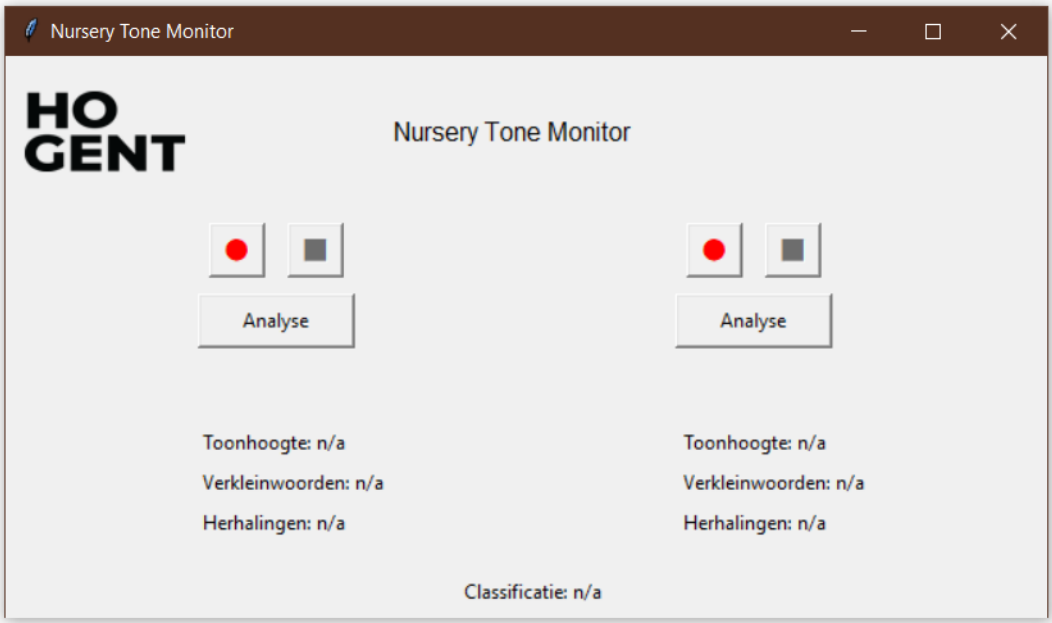
3.0.3 Eigenschappen Praat vergelijken met python bibliotheken

Nadien werd er een python script opgesteld om de opnames te verwerken op toonhoogte aan de hand van python bibliotheken. Als eerste werd hiervoor de ‘CREPE pitch tracker‘ gebruikt. Deze bibliotheek gaf waarden van de toonhoogte die gelijkaardig waren aan die van de Praat-software. De toonhoogten van de ‘CREPE pitch tracker‘ weken echter wel af van deze die ‘Praat‘ berekend had. De afwijkingen bleven echter binnen 5 procent. De verwerkingstijd van de opnames was zeer traag en bleek niet praktisch te zijn in reële situaties. Zo duurde het ongeveer 40 seconden om een opname met een lengte van anderhalve minuut te analyseren. Vervolgens werd er gebruikt gemaakt van ‘Parselmouth‘. Deze bibliotheek gebruikt de Praat source-code en gaf dus gelijke waarden als Praat zelf. De verwerkingstijd voor een opname van gelijke lengte bedroeg enkele seconden tegenover de 40 seconden die ‘CREPE‘ nodig had, daarom werd besloten met deze bibliotheek verder te gaan in het onderzoek.

3.0.4 Praktisch gebruik Nursery Tone Monitor

Om een indicatie te krijgen van hoe de ‘Nursery Tone Monitor‘ praktisch gebruikt kan worden, werd er een kleine applicatie gemaakt in python. Voor de user interface van deze applicatie werd de ‘Tkinter‘ bibliotheek voor python gebruikt. ‘Tkinter‘ is de standaard Graphical User Interface (GUI) van python en is zeer toegankelijk. Het laat toe python scripts snel in een GUI te verwerken wat de ‘Nursery Tone Monitor‘ in dit geval praktischer zal maken in gebruik.

Deze applicatie laat het toe twee opnames te maken en telkens deze opnames te analyseren op de gemiddelde toonhoogte. De analyse wordt gedaan aan de hand van ‘Parselmouth‘. Op deze manier kan de toonhoogte makkelijk en snel gecontroleerd worden in de twee opnames. De testapplicatie toont ook het aantal verkleinwoorden en zinnen die herhaald worden om aan te geven dat dit in de toekomst toegevoegd kan worden. Dit is echter louter illustratief en heeft in deze testapplicatie geen functie. Figuur 3.1 toont een voorbeeld van de lay-out van de Nursery Tone Monitor test applicatie.



Figuur 3.1: Nursery Tone Monitor test app.

3.0.5 CNN model

Als laatste werd ook geprobeerd elderspeak te detecteren aan de hand van een convolutioneel neuraal netwerk (CNN). De opnames werden gestructureerd en verzameld op 'Google Drive'. De opnames werden ingelezen en omgezet naar normale en mel-frequentie spectrogrammen. De mel-frequentie spectrogrammen werden vervolgens omgezet naar Numpy-arrays en werden verdeeld in test- en train-data. Oorspronkelijk werd 20 procent van de input data gebruikt als testdata, maar na training werd dit verhoogd tot 40 procent. De reden hiervoor was dat het aantal audiosamples in de testdata te klein was en er soms hoge nauwkeurigheden (van 70 procent en meer) verkregen werden omdat het model 'toevallig' enkele keren een goede classificatie gemaakt had. Door de testdata groter te maken werd deze toevalligheid tegen gegaan. Deze inputdata werd gebruikt om een model te trainen en zo hopelijk tot een performant model te bekomen die correct elderspeak kan detecteren uit een bepaalde audiofile.

3.0.6 Data augmentation

Aangezien de dataset zeer beperkt was (20 personen), was het nodig deze data uit te breiden aan de hand van data augmentatie. De transformaties die gebruikt werden in dit onderzoek waren: 'het verplaatsen van het spectrogram in de tijd en het toevoegen van ruis'. In dit onderzoek werden geen transformaties uitgevoerd die invloed hadden op de toonhoogte aangezien dit een te belangrijke factor was voor de detectie van elderspeak. Door iedere opname te voorzien van gerandomiseerde ruis en de opname te verplaatsen op de tijds-as was het mogelijk de dataset te verdrievoudigen. Het zou mogelijk moe-

ten zijn de dataset verder uit te bereiden door het combineren van deze transformaties op de dataset, maar dit is in deze bachelorproef niet gedaan. Bij het gebruiken van de geaugmenteerde data werd 20 procent van de inputdata gebruikt als testdata.

4. Resultaten

4.0.1 Opnames

Tabel 4.1 toont de gemiddelde toonhoogte voor iedere deelnemer. In de tabel staan steeds de gemiddelde waarden van de toonhoogte per casus. De casus waar de toonhoogte gemiddeld de hoogste waarde had, staat vetgedrukt. Bij de analyse van deze gemiddeldes werden frequenties lager dan 75 Hz en hoger dan 600 Hz niet meegenomen. In deze resultaten is het zichtbaar dat casus 2 en casus 3 gemiddeld een hogere toonhoogte hadden dan casus 1. Dit waren eveneens de casussen waar elderspeak van toepassing was.

4.0.2 Parselmouth

De Parselmouth bibliotheek maakt het mogelijk om via python enkele functies van Praat aan te roepen. Aangezien deze bibliotheek gebruikt maakt van de Praat toepassing, worden dezelfde waarden verkregen voor de toonhoogte. De verwerkingstijd van de analyse van een audiobestand duurt slechts enkele seconden wat het zeer geschikt maakt om in de praktijk te gebruiken. De testapplicatie is makkelijk in gebruik en kan mits eventuele toevoegingen en aanpassing gebruikt worden in reële situaties.

4.0.3 CNN

Het convolutioneel neuraal netwerk (CNN) behaalde slechte resultaten. De accuraatheid van het model schommelt tussen de 45 en 60 procent. Bij het trainen van het CNN werden meermaals problemen ontdekt door een tekort aan RAM-geheugen. Hierdoor was het niet mogelijk eventuele extra lagen toe te voegen. Aangezien er gewerkt werd met een kleine

Persoon	Casus 1 (Hz)	Casus 2 (Hz)	Casus 3 (Hz)
1	130.25	**137.06**	133.39
2	140.15	**187.64**	178.37
3	207.46	**234.60**	233.89
4	223.69	257.29	**269.06**
5	110.75	111.40	**120.39**
6	148.19	176.57	**199.67**
7	209.13	236.51	**254.38**
8	222.27	**268.42**	261.56
9	89.37	100.29	**146.69**
10	119.95	140.28	**153.34**
11	208.73	263.35	**265.30**
12	210.28	213.10	**218.37**
13	208.51	**243.98**	225.79
14	200.95	**290.52**	289.24
15	192.97	216.43	**222.29**
16	216.38	**266.09**	254.25
17	196.22	189.81	**216.98**
18	206.66	228.72	**237.12**
19	133.07	150.61	**160.19**
20	137.47	173.27	**180.48**

Tabel 4.1: Gemiddelde toonhoogte per persoon per casus volgens Praat. De hoogste waarde per casus staat vetgedrukt.

dataset die ook gesplitst werd in test- en train-data, was er zeer snel sprake van overfitting. Zo maakt het model een goede voorspelling wanneer het een audiosample krijgt die deel was van de testdata, maar is dit niet perse het geval zijn wanneer nieuwe data gepresenteerd wordt. Aangezien de classificatie bestaat uit slechts 2 klassen, is het zeer moeilijk dit model te beoordelen. Dit omdat statistisch gezien een correcte classificatie gegokt kan worden met een accuraatheid van 50 procent. Dit zou ook een reden kunnen zijn voor de huidige 40-60 procent accuraatheid van het model.

4.0.4 CNN met data augmentation

De augmentatie van de dataset had in dit geval geen effect op de accuraatheid van het model. De accuraatheid bleef schommelen tussen de 40-60 procent. Deze resultaten met het gegeven dat er een accuraatheid van 50 procent behaald kan worden door middel van gokken, maakt dat dit model in zijn huidige staat geen meerwaarde biedt in de detectie van elderspeak.

5. Conclusie

Inleiding

Er zijn verschillende manieren mogelijk om elderspeak te detecteren. Deze bachelorproef tracht elderspeak te detecteren aan de hand van de gemiddelde toonhoogte en trachtte door middel van een convolutioneel neuraal netwerk een correcte classificatie van elderspeak te maken. Hieronder een korte bespreking van de bevindingen van deze bachelorproef.

5.1 Toonhoogte detecteren aan de hand van python applicatie

Het detecteren van de toonhoogte blijkt een zeer goede manier om elderspeak te detecteren. Het gebruik van een python script en python bibliotheken behaald zeer goede resultaten en kan makkelijk aangepast worden. Deze bachelorproef werkt momenteel met twee opnames die manueel gestart moeten worden. Een mogelijke aanpassing van deze applicatie zou kunnen zijn dat er een continue detectie is, die op het einde van de opname een volledig overzicht kan generen van de toonhoogte. Zulk overzicht zou kunnen aantonen waar en wanneer elderspeak gebruikt werd, om de spreker in kwestie zo persoonlijk mogelijke feedback te kunnen geven op het gebruik hiervan.

De python applicatie kan ook uitgebreid worden door te gaan detecteren op het aantal herhaalde woorden en/of zinnen alsook het gebruik van verkleinwoorden. Wanneer deze zaken gecombineerd kunnen worden, zou een praktische toepassing van de ‘Nursery Tone Monitor‘ snel in de praktijk toegepast kunnen worden.

5.2 Convolutioneel neuraal netwerk

Deze bachelorproef heeft geen meerwaarde kunnen aantonen voor het gebruik van een CNN. Wegens omstandigheden was het niet mogelijk te beschikken over een grote dataset wat leidt tot slechte voorspellingen van het model. Het model voorspelt een classificatie aan dezelfde accuraatheid dan dat gokken zou teweegbrengen. Dit maakt het huidige model onbruikbaar in de praktijk. Naar de toekomst toe zou dit model op punt gebracht kunnen worden door gebruik te maken van een grotere dataset. Wanneer deze geaugmenteerd kan worden, zou het mogelijk moeten zijn een degelijke voorspelling te maken. Hierbij moet opgemerkt worden dat er voldoende RAM moet zijn om het model te trainen aangezien deze bachelorproef hier enkele problemen mee ondervonden heeft. Wanneer een voldoende grote dataset gebruikt wordt, zou er hyperparameter tuning (zodanig aanpassen van bepaalde parameters van het model) ervoor kunnen zorgen dat een CNN gebruikt kan worden voor de detectie van elderspeak.

Ideaal zouden beide methoden gecombineerd kunnen worden, waar manueel aan de hand van de toonhoogte, herhalende woorden/zinnen en verkleinwoorden en automatisch aan de hand van een CNN een correcte detectie kan plaatsvinden van elderspeak.

5.3 Toekomst van dit project

Deze ‘Nursery Tone Monitor‘ zou zeker uitgewerkt kunnen worden tot een praktisch gegeven. Hiervoor stelt deze bachelorproef voor toekomstig onderzoek te verrichten op het praktische gebruik van de applicatie en welke verbeteringen eventueel mogelijk kunnen zijn. Deze verbeteringen kunnen zoals hierboven aangegeven te maken hebben met de manier waarop de toonhoogte geanalyseerd wordt (manueel of automatisch). Er zou gelijktijdig aan deze bachelorproef een medestudent onderzoek verrichten naar de detectie van verkleinwoorden en herhaalde zinnen. De combinatie van deze twee onderzoeken zou kunnen zorgen voor een mooie meerwaarde betreft de ‘Nursery Tone Monitor.‘

Ondanks dat dit onderzoek geen meerwaarde kunnen aantonen heeft voor het gebruiken van een convolutioneel neuraal netwerk voor de detectie van elderspeak, stelt deze bachelorproef voor deze manier van detectie niet af te schrijven. Het detecteren van elderspeak aan de hand van een CNN blijft een interessante denkpiste die zeker verder onderzocht moet worden. Hetzij met een grotere en meer variërende dataset.

A. Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

A.1 Introductie

A.1.1 Context

Wereldwijd neemt het aandeel ouderen in de bevolking spectaculair toe. In 1990 was 9,2 procent van de wereldbevolking zestig jaar of ouder, terwijl verwacht wordt dat tegen 2050 maar liefst 21,1 procent van de wereldbevolking minstens zestig jaar zal zijn. Ook in Vlaanderen wordt deze trend gevolgd. In 2000 was het aandeel van zestigplussers 16,7 procent. In 2020 was dit aandeel met 3,8 procent gestegen tot 20,5 procent. Volgens de huidige statistiek lijkt deze trend zich naar de toekomst toe ook door te zetten (Bynens, 2020). In deze vergrijzende context, komen zorgverleners meer en meer in contact met deze oudere zorgvragers. Dit geldt zowel voor zorgverleners die werkzaam zijn in woonzorgcentra, maar evengoed voor zorgverleners die werken binnen een thuiscontext.

A.1.2 Nood

Aangezien een goed sociaal contact een positieve invloed heeft op de mentale gezondheid van een individu, is het belangrijk dat de behoefte aan sociaal contact bij de ouderen wordt vervuld. Toch wordt deze manier van communiceren vaak als ontoereikend beschouwd door zowel de zorgvrager als zorgverlener (Balsis & Carpenter, 2006). Hoewel het ge-

bruik van elderspeak door een beperkte groep ouderen als positief kan worden ervaren, wordt het door de meerderheid eerder als negatief beschouwd. Zo wordt het vaak ervaren als onrespectvol en vinden de ouderen de instructies die gegeven worden in elderspeak verwarrend of onduidelijk. Verder wordt de communicatie door deze ouderen ervaren als moeizaam (Lowery, 2013).

A.1.3 Nursery Tone Monitor

Om zorgverleners bewust te maken of ze (al dan niet bewust) deze communicatiestijl gebruiken, wil HoGent een applicatie ontwikkelen die real time feedback kan geven over de spreekstijl van een zorgverlener in opleiding. Deze Nursery Tone Monitor zal in real time aantonen of er bepaalde kenmerken van elderspeak in een gesprek aanwezig zijn om zo snel en effectief feedback te kunnen geven aan de zorgverlener. Deze bachelorproef zal zich focussen op het eerste deel van de ontwikkeling van deze applicatie en zal onderzoeken wat de beste manier is om elderspeak softwarematig te detecteren. Het resultaat kan gebruikt worden om de Nursery Tone Monitor verder uit te werken tot een functioneel gegeven.

A.2 State-of-the-art

A.2.1 Elderspeak

Elderspeak is een vorm van communicatie waarbij de spreker gebruikt maakt van een zeer simpel taalgebruik. Dit kenmerkt zich door het gebruik van een simpele grammatica en beperkte woordenschat. Andere belangrijke factoren zijn een verhoogde toonhoogte en volume tijdens het spreken. Elderspeak wordt voornamelijk gebruikt door jongere volwassenen tegenover oudere volwassenen (Kemper, FinterUrczyk e.a., 1998). De reden hiervoor ligt bij het stereotype waarbij de oudere wordt bestempeld als minder begaafd en minder snel van begrip. Dit stereotype wordt versterkt naargelang de oudere extra zorg nodig heeft of leidt aan neurologische aandoeningen zoals dementie (Balsis & Carpenter, 2006).

A.2.2 Toonhoogte

Aangezien een verhoogde toonhoogte een belangrijke factor is van elderspeak, lijkt het verstandig om deze te detecteren. Voor het detecteren van de toonhoogte (of F0 detectie om meer precies te zijn) zijn verschillende algoritmes ontwikkeld, de meest gebruikte hiervoor zijn: Praat, RAPT, STRAIGHT, YIN en SWIPE (Strömbergsson, 2016). Het Praat algoritme wordt het meest frequent gebruikt aangezien dit momenteel de beste resultaten biedt op het detecteren van de toonhoogte. Een andere studie die de algoritmes vergelijkt op het verwerken van geluidsfragmenten met een zeker ruis (Jouvet & Laprie, 2017), toonde gelijkaardige resultaten. Beide studies tonen aan dat het Praat algorimte het minste fouten maakt in het detecteren van de toonhoogte. Dit algoritme lijkt dus de

logische keuze om mee verder te werken in dit onderzoek.

A.2.3 Detectie

Voor het softwarematig detecteren van de toonhoogte bestaan er reeds enkele mogelijkheden. Het programma Praat kan gebruikt worden voor het analyseren en visualiseren van verschillende aspecten van een audiobestand. Dit programma maakt gebruik van het Praat-algoritme wat er voor ons reeds veelbelovend uitzag aangezien het ook goed was in omgevingen met ruis (Jouvet & Laprie, 2017). Het programma is ideaal om na te gaan waar de verschillen liggen in audiofragmenten met en zonder elderspeak. Een andere manier om toonhoogte te kunnen detecteren is door middel van de CREPE Pitch Tracker voor python. Deze bibliotheek maakt gebruik van diepe convolutionele neurale netwerken die schattingen kunnen maken van de toonhoogte in bepaalde audiofragmenten. CREPE werd vergeleken met het SWIPE en YIN algoritme en behaalde betere resultaten op zowel de detectie van de toonhoogte, als de detectie van de toonhoogte wanneer er achtergrond geluid of een bepaalde ruis aanwezig is (Kim e.a., 2018). Dit is belangrijk aangezien de Nursery Tone Monitor uiteindelijk gebruikt zal worden op plaatsen waar deze ruis aanwezig is.

A.3 Methodologie

In dit onderzoek zal eerst onderzocht worden wat de verschillen zijn tussen gewone spraak en elderspeak en of deze zichtbaar zijn in een spectrogram. Er zullen audiofragmenten verzameld worden van gesprekken met en zonder elderspeak. Deze audiofragmenten zullen door middel van de Praat software geanalyseerd worden om na te gaan of er verschillen in toonhoogte en volume detecteerbaar zijn. Nadien zal CREPE Pitch Tracker voor python gebruikt worden om na te gaan of de verschillen die gevonden worden door CREPE vergelijkbaar zijn met deze van de Praat software.

A.4 Verwachte resultaten

Er wordt verwacht dat het gemiddelde volume en de gemiddelde toonhoogte hoger zullen liggen in fragmenten waar elderspeak toegepast of uitgelokt werd. Er wordt verwacht dat de Praat software deze verschillen duidelijk zal kunnen aantonen en dat ook de CREPE Pitch Tracker in staat zal zijn deze verschillen te detecteren.

A.5 Verwachte conclusies

Uit dit onderzoek zal duidelijk worden op welke manier elderspeak het best gedetecteerd kan worden in een audiofragment. Er wordt verwacht dat dit mogelijk zal zijn op basis

van het volume en de toonhoogte van een gesprek aangezien dit twee belangrijke kenmerken zijn van elderspeak. Omdat AI en machine-learning algoritmes zeer goed zijn in het herkennen van patronen, zou het mogelijk moeten zijn dit toe te passen op gesproken audiofragmenten. Aangezien python vaak gebruikt wordt voor dit soort toepassingen en het meestal ook goed gedocumenteerd is, wordt er ook verwacht dat elderspeak detecteerbaar via deze taal en zou zo een praktische implementatie van de Nursery Tone Monitor mogelijk moeten zijn.

B. Github Repository

Het formulier waarin opnames gevraagd werden, alsook codefragmenten met betrekking tot de python applicatie en het convolutioneel neuraal netwerk kunnen terug gevonden worden op de Github Repository (Beeckman, 2021) .

-

Bibliografie

Atria, J. J. (2015, januari 15). *Pitch in Praat*. https://www.pinguinorodriguez.cl/blog/pitch-in-praat/

Balsis, S. & Carpenter, B. D. (2006). Evaluations of Elderspeak in a Caregiving Context. *Clinical Gerontologist*, *29*(1), 79–96. https://doi.org/10.1300/j018v29n01_07

Beeckman, G. (2021, mei 27). *Github Repository met codefragmenten*. https://github.com/GlennBeeckman/BachelorProefTIHoGent

Boersma, P. (1993). Accurate Short-Term Analysis Of The Fundamental Frequency And The Harmonics-To-Noise Ratio Of A Sampled Sound.

Bynens, J. (2020, juli 14). *Bevolking naar leeftijd en geslacht*. https://www.statistiekvlaanderen.be/nl/bevolking-naar-leeftijd-en-geslacht

Chen, J. (2020, december 23). *Neural Network* (M. Boyle, Red.). https://www.investopedia.com/terms/n/neuralnetwork.asp

Francart, T. & Eneman, K. (2008). Analyse van de zangstem: geluiden in beeld.

Hart, J. (2011, oktober 1). *Musical Notes Explained Simply*. http://theprovincialscientist.com/?p=1576

IBM Cloud Education. (2020, augustus 17). *Neural Networks*. https://www.ibm.com/cloud/learn/neural-networks

Jadoul, Y., Thompson, B. & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Jouvet, D. & Laprie, Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. *2017 25th European Signal Processing Conference (EUSIPCO)*. https://doi.org/10.23919/eusipco.2017.8081482

Kamp, J. V. (2020, mei 14). *What is a spectrogram?* https://vibrationresearch.com/blog/what-is-a-spectrogram/

Kemper, S., Finter-Urczyk, A., Ferrell, P., Harden, T. & Billington, C. (1998). Using elderspeak with older adults. *Discourse Processes*, *25*(1), 55–73. https://doi.org/10.1080/01638539809545020

Kemper, S., FinterUrczyk, A., Ferrell, P., Harden, T. & Billington, C. (1998). Using elderspeak with older adults. *Discourse Processes*, *25*(1), 55–73. https://doi.org/10.1080/01638539809545020

Kim, J. W., Salamon, J., Li, P. & Bello, J. P. (2018). Crepe: A Convolutional Representation for Pitch Estimation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2018.8461329

Lowery, M. A. (2013, mei 5). *Elderspeak: Helpfull or harmfull? A systematic review of speech to eldery adults*. Graduate Faculty of Auburn University. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.910.8093&rep=rep1&type=pdf

Nanni, L., Maguolo, G., Brahnam, S. & Paci, M. (2021). An Ensemble of Convolutional Neural Networks for Audio Classification.

Ramaseshan, A. (2013). *Application of Multiway Methods for Dimensionality Reduction to Music* (proefschrift).

Saha, S. (2018, december 15). *A Comprehensive Guide to Convolutional Neural Networds - the ELI5 way*. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Salamon, J. & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/lsp.2017.2657381

Shorten, C. & Khoshgoftaar, T. M. (2019). *A survey on Image Data Augmentation for Deep Learning*. https://doi.org/10.1186/s40537-019-0197-0

Strömbergsson, S. (2016). Today's Most Frequently Used F0 Estimation Methods, and Their Accuracy in Estimating Male and Female Pitch in Clean Speech. *Interspeech 2016*. https://doi.org/10.21437/interspeech.2016-240

The Physics Classroom. (g.d.). *Pitch and Frequency*. https://www.physicsclassroom.com/class/sound/Lesson-2/Pitch-and-Frequency

van Lieshout, P. (2003). PRAAT: Short Tutorial.