

Verbeteren van VR stottertherapie door het herkennen en verwerken van verbale antwoorden met spraakherkenningssoftware.

Jona De Neve.

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

Promotor: Mevr. Lena De Mol

Co-promotor: Mevr. Jana Van Damme

Academiejaar: 2022–2023

Tweede examenperiode

Departement IT en Digitale Innovatie .

**HO
GENT**

Woord vooraf

Deze bachelorproef werd geschreven in het kader van het voltooien van de opleiding Toegepaste Informatica afstudeerrichting Mobile & Enterprise development. Hoewel dit onderwerp niet valt onder mijn afstudeerrichting, koos ik ervoor uit interesse van de huidige stand van Artificiële Intelligentie en hoe het verder evolueert in de toekomst. Zo volg ik de vooruitgang ook wekelijks. Daarnaast wil ik graag een aantal mensen bedanken, zonder wie deze proef niet tot stand zou zijn gekomen. In de eerste plaats wil ik mijn promotor, Lena De Mol, bedanken voor de begeleiding doorheen mijn bachelorproef.

Daarnaast wil ik mijn co-promotor, Jana Van Damme, bedanken voor de hulp tijdens de opzet van deze bachelorproef en voor alle nuttige feedback onderweg.

Samenvatting

Inhoudsopgave

Lijst van figuren	vii
1 Inleiding	1
1.1 Probleemstelling	2
1.2 Onderzoeksvraag	2
1.3 Onderzoeksdoelstelling	2
1.4 Opzet van deze bachelorproef	2
2 Stand van zaken	3
2.1 Stotteren	3
2.2 Virtuele realiteit	4
2.3 Artificiële intelligentie	4
2.3.1 Spraakherkenning	5
2.3.2 Begrijpen van natuurlijke taal	6
3 Long list	8
3.1 Spraakherkenning	8
3.1.1 Google Cloud Speech-to-text API	8
3.1.2 Amazon Transcribe	9
3.1.3 Microsoft Azure Speech Services	10
3.1.4 Mozilla DeepSpeech	11
3.1.5 Whisper	12
3.1.6 Kaldi ASR	12
3.2 Natuurlijke taalverwerking	14
3.2.1 GPT series	14
3.2.2 BERT	14
3.2.3 FlairNLP	14
3.2.4 RoBERTa	15
4 Methodologie	16
4.1 Requirementsanalyse	16
4.1.1 Functionele requirements	16
4.1.2 Niet-functionele requirements	18
4.1.3 MoSCoW	18
4.1.4 Use Case	19
4.2 Proof-of-concept	19

5 Conclusie	20
A Onderzoeksvoorstel	21
A.1 Introductie	21
A.2 State-of-the-art	22
A.3 Methodologie	23
A.4 Verwacht resultaat, conclusie	24
Bibliografie	27

Lijst van figuren

2.1	Dimensions in Testimony - Systeem Architectuur (Traum e.a., 2015) . . .	5
2.2	NPCEditor - Systeem Architectuur (Leuski & Traum, 2010)	7
3.1	Prijzen Amazon Transcribe voor Parijse server (Amazon, 2023)	10
A.1	Dimensions in Testimony - Systeem Architectuur (Traum e.a., 2015) . . .	22
A.2	NPCEditor - Systeem Architectuur (Leuski & Traum, 2010)	25
A.3	Flowchart methodologie	26

1

Inleiding

In de reclamespot 'The Impact Will Be Real' over de Metaverse toont Meta ([2022](#)) hun visie over de rol die Virtual Reality (VR) speelt in de toekomst. Zo laten ze verschillende toepassingen ervan in het onderwijs zien. Jammer genoeg staat onze technologie nog niet zo ver als wat er in de reclame gezien kan worden maar ook nu al vindt VR zich een baan in verschillende opleidingen.

De Hogeschool van Gent maakt ook gebruik van VR om studenten de kans te geven in meer realistische situaties te oefenen. Hiervoor zijn twee verschillende technieken gebruikt. Allereerst heb je het renderen van een omgeving. Dit laat de gebruiker een interactieve wereld van 3D modellen ontdekken. Zo bestaan er drie virtuele kamers waarin de student kan oefenen. De andere manier is aan de hand van een 360° opname die wordt gemaakt aan de hand van een 360° camera. Omdat dit een opname van de werkelijkheid neemt, ziet deze methode er realistischer uit.

Dit is waar er op het probleem wordt gestoten. Aangezien de tweede methode werkt met een opname moet er naar verschillende fragmenten gesprongen worden naargelang het antwoord dat de gebruiker ingeeft. Dit wordt handmatig gedaan door een begeleider. Hierdoor staat de oefening tijdelijk stil wat de echtheid van de situatie weghaalt. Daarom wordt in deze bachelorproef toegepaste informatica onderzocht hoe het overschakelen anders kan aangepakt worden zodat het voor de gebruiker realistischer aanvoelt. Hiervoor kijken we richting Artificiële Intelligentie (AI).

Het afgelopen jaar is de populariteit van AI enorm gestegen. Met text-to-image models zoals DALL-E 2, Imagen en Stable diffusion die een gegeven tekst prompt kunnen omzetten in afbeeldingen,

1.1. Probleemstelling

Wanneer de stotterpatiënt antwoordt op de gestelde vraag wordt er niet automatisch overgeschakeld naar een volgend fragment. Er moet namelijk handmatig geklikt worden op het gewenste hoofdstuk door iemand die de oefening kent. Dit zorgt dat er steeds een begeleider aanwezig moet zijn. Ook zal de dynamiek van de oefening verbroken worden omdat het pas verder kan gaan wanneer de begeleider het juiste fragment vindt.

1.2. Onderzoeksvraag

Om dit probleem te verhelpen wordt hiervoor gekeken hoe AI ons kan te hulp schieten om stotterpatiënten een meer interactieve ervaring geven. Om dit te realiseren zullen de volgende vragen beantwoord worden:

- Hoe gaat de applicatie registreren wat er werd gezegd?
- Wat kan er mis lopen bij het registreren van spraak?
- Hoe kiest de applicatie een volgend fragment op basis van de gegenereerde tekst?

1.3. Onderzoeksdoelstelling

In deze bachelorproef word een proof of concept opgesteld om te kijken hoe zo een AI-applicatie te werk zou kunnen gaan. De bedoeling van deze applicatie is eerst en vooral registreert wanneer de gebruiker aan het antwoorden is. Vervolgens transcribeert het wat er gezegd wordt. Als laatste gaat het op basis van de context een gepast fragment zoeken om verder mee te gaan.

1.4. Opzet van deze bachelorproef

De rest van deze bachelorproef is als volgt opgebouwd:

In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.

In Hoofdstuk 4 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

In Hoofdstuk 5, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2

Stand van zaken

Zoals vermeld in de inleiding heeft het zorglab van Hogeschool Gent verschillende doeleinden. Zo kunnen studenten geneeskunde leren hoe ze een operatie moeten uitvoeren zonder iemand als testpersoon te gebruiken. Anderzijds kan het een persoon die plankenkoorts heeft een presentatie of een toespraak leren houden. In onze situatie zal een patiënt met een stotter leren een vlot gesprek te houden zoals bijvoorbeeld wanneer hij op sollicitatie gaat. Dit wordt gerealiseerd aan de hand van 360° opnames die voor de patiënt opnieuw afgespeeld kunnen worden in VR.

2.1. Stotteren

Stotteren is een spraakstoornis dat gekenmerkt wordt door onderbrekingen en verstoringen in het vloeiend spreken. De aandoening komt vaker voor bij kinderen en gaat bij 80% over naarmate ze opgroeien (Gordon, 2002). Er zijn verschillende types van onvloeiendheid in spraak (Chee e.a., 2009): tussenwerpsels (woorden zoals 'uh' of 'eh'), herzieningen (het aanpassen van de inhoud of de grammaticale structuur van de zin), onafgemaakte zinnen, herhaling, aanhoudende geluiden en afgebroken woorden. Een stotter kan voorkomen in verschillende mate, sommigen zullen meer last hebben zichzelf te uiten terwijl bij anderen de stoornis niet hard opvalt. De specifieke oorzaak van een stotter is nog niet doorgrond. Uit het onderzoek van Gordon blijkt echter dat de rechter- en linkerhersenhelft verschillende en tegenovergestelde rollen lijken te spelen bij het ontstaan van stotteren. Symptomen van stotteren worden geassocieerd met activering van de voorste gebieden in de linkerhersenhelft, terwijl zowel voorste als achterste perisylviaanse gebieden in de rechterhersenhelft geactiveerd worden wanneer het spreken vloeiender wordt. Deze verbetering kan mogelijk te wijten zijn aan de koppeling van motorische en zintuiglijke gebieden in de rechterhersenhelft.

2.2. Virtuele realiteit

Hoewel Virtual Reality tegenwoordig veel gevorderd is, bestaat het al voor een lange tijd. Het eerste toestel dat de werkelijkheid nabootste was Morton Heilig's Sensorama uit 1962. Dit liet de gebruiker ervaren hoe het voelde om op een motor door Boston te rijden. De opvolgende jaren werden ook andere toestellen uitgevonden zoals de 'Sword of Damocles' die de locatie van het hoofd en de ogen volgde en Nintendo's 'power glove' voor de NES (Boas, 2012).

De eerste vermelding van de term VR daarentegen kwam pas rond 1985 zegt Bryson (2013). Hij verteld over hoe Jaron Lanier de term 'virtual reality' gebruikte om het Virtual Interactive Environment Workstation (VIEW) lab van NASA te beschrijven. De daaropvolgende jaren werd de term steeds populairder met als gevolg dat het ruim werd toegepast op verschillende toepassingen. Daarom moest de term VR goed gedefinieerd worden:

Virtual Reality is the use of computer technology to create the effect of an interactive three-dimensional world in which the objects have a sense of spatial presence. (Bryson, 2013)

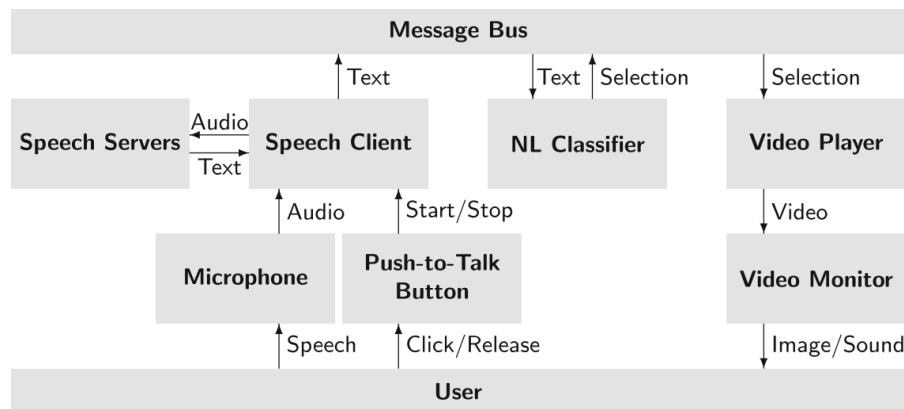
Tegenwoordig zijn VR toestellen te vinden in vele soorten en maten. Zo zijn er headsets gemaakt voor PC of consoles en anderen voor smartphones. Dankzij software en hardware verbeteringen bestaan er nu zelfs headsets die autonoom werken. Dit zorgt ervoor dat kabels en complexe set-ups niet nodig zijn. Deze verbeteringen zorgen ook dat de digitale wereld realistischer doet aanvoelen.

2.3. Artificiële intelligentie

Het afgelopen jaar is de populariteit van AI enorm gestegen. Met text-to-image models zoals DALL-E 2, Imagen en Stable diffusion die een gegeven tekst prompt kunnen omzetten in afbeeldingen. Daarnaast zijn er ook applicaties die language models gebruiken zoals Chat-GPT, spraak assistenten en Github Copilot. Zij kunnen teksten verwerken en accurate antwoorden genereren op basis van hun kennis ook al zijn deze niet altijd accuraat (Meer hierover in 2.3.2)

Om het overschakelen naar een volgend fragment te laten gebeuren zonder begeleiding en zonder veel tijd te verliezen, zouden verbale antwoorden gegeven kunnen worden. Een mooi voorbeeld van een gelijkaardige interactie is het project 'Dimensions in Testimony' van de USC Shoah Foundation (2020). Daar kunnen bezoekers vragen stellen aan een overlevende van de Holocaust die op voorhand een volledig interview heeft afgelegd. Dit werd mogelijk gemaakt dankzij het systeem dat erachter zit (Figuur 2.1). Dit systeem is opgebouwd uit volgende componenten: een software voor spraakherkenning (ASR) die de gebruikers verbale vraag in tekst omzet, een Natural Language Classifier (NLC) die op basis van de gegenereerde tekst een antwoord via een audio/video fragment voorziet en een mediaspeler die

de fragmenten kan afspelen met tussendoor een inactieve animatie (Traum e.a., 2015).



Figuur (2.1)

Dimensions in Testimony - Systeem Architectuur (Traum e.a., 2015)

Een ander voorbeeld zijn de home assistenten zoals Google Home, Amazon Echo of Apple HomePod. Zij staan op stand-by tot de gebruiker het triggerwoord zegt, 'Hey Google' bijvoorbeeld. Eenmaal geactiveerd luisteren ze naar wat de gebruiker zegt en zetten ze dit aan de hand van een ASR om naar tekst. Daarna haalt het met natuurlijke taalverwerking (NLP) de intentie en sleutelwoorden uit de tekst en genereert daarmee een gepast antwoord. Als laatste zet het gegenereerde tekst om naar spraak.

2.3.1. Spraakherkenning

Net als VR is spraakherkenning de laatste jaren veel vooruitgegaan, dit komt omdat er meer data beschikbaar is, de computers sneller en de algoritmes, genaamd deep neural networks, beter zijn. Deze netwerken worden getraind op grote datasets bestaande uit diverse spraakfragmenten. Hieruit leert het de patronen en kenmerken van menselijke spraak. Maar perfect zullen ze nooit zijn zegt van Hessen (2020). Dit is geen verrassing aangezien zelf mensen vaak nog problemen hebben bij het verstaan van een andere. Er zijn verschillende stappen waar de ASR de fout kan ingaan.

Domein problemen

Onder de domein problemen valt ten eerste galm en geluidshinder. Dit komt voor wanneer een opname gemaakt wordt in een omgeving met veel achtergrond lawaai. Net als mensen heeft de ASR moeilijkheden met het verstaan van de spreker wanneer deze overstemd wordt door de omgeving (Alharbi e.a., 2021). Naast galm en geluidshinder is het ook moeilijker wanneer twee of meer mensen door elkaar spreken. Dit heet spraak overlapping. Een derde domein probleem is domain mismatch. Dit houdt in dat er een discrepantie is tussen het domein gebruikt voor het

model te trainen en de use case waar het model in wordt gebruikt. Zo zijn bijvoorbeeld nieuwe accenten of andere omgevingen onbekend voor het model. Daarnaast kan het doeleinde een ander jargon bevatten wat leidt tot out-of-vocabulary fouten.

Natuurlijke taalverwerking

Zoals vermeld in de vorige paragraaf is de use case van groot belang. Taal dat gebruikt wordt verschilt op vele manieren: jongeren gaan andere woorden gebruiken dan hun senioren, iemand in de IT sector gaat een andere woordenschat hebben dan een dokter. Deze verschillen kunnen leiden tot out-of-vocabulary fouten. Dit houdt in dat een bepaald model niet getraind is op bepaalde manier van spreken of woordenschat. Een andere oorzaak van OOV is het trainen van een model op een te kleine dataset.

Ook niet iedereen spreekt woorden op dezelfde manier uit. Afhankelijk van de sprekers afkomst zullen ze een bepaalde taal met een bepaald dialect spreken. Ook wanneer ze in een taal anders dan hun moedertaal praten zal er vaak sporen van hun eigen taal herkenbaar blijven. Dit zorgt ervoor dat de uitspraak van een woord niet altijd hetzelfde is.

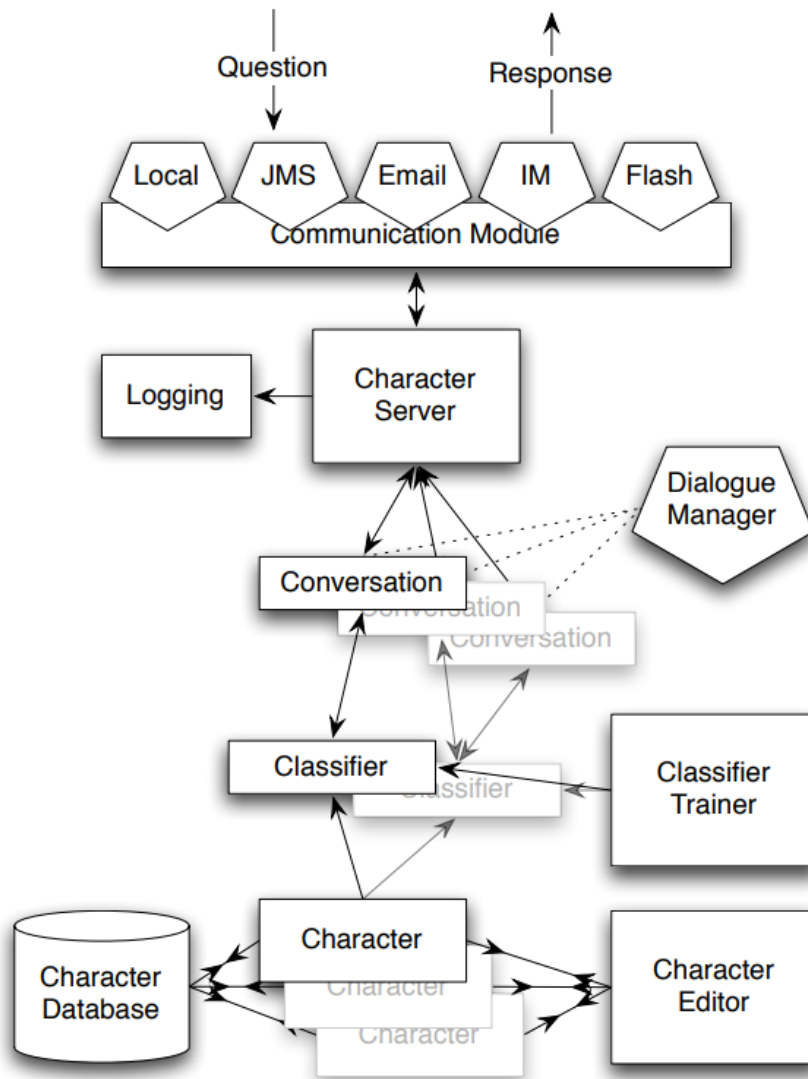
Efficiëntie van apparatuur

Als laatste hebben we de gebruikte toestellen. De kwaliteit van de opnameapparatuur kan de nauwkeurigheid van de ASR beïnvloeden. Het gebruik maken van een high-end microfoon tegenover een goedkope kan betere resultaten creëren.

2.3.2. Begrijpen van natuurlijke taal

Naast de ASR hebben we ook het begrijpen van natuurlijke taal (NLU). In 'Dimension in Testimony' maakten ze hiervoor gebruik van NCPeditor (Figuur 2.2). De classifier in dit systeem berekent welke antwoorden op de ingegeven tekst kunnen gegeven worden door de taalmodellen van beide te vergelijken en de antwoorden te rangschikken. Zolang de ingegeven tekst een foutmarge lager dan 50% blijft, zal het antwoord ongewijzigd blijven (Leuski & Traum, 2010).

Een andere bekend taalmodel is GPT-3. Dit model ligt aan de basis van de populaire chatbot ChatGPT.

**Figuur (2.2)**

NPCEditor - Systeem Architectuur (Leuski & Traum, 2010)

3

Long list

3.1. Spraakherkenning

3.1.1. Google Cloud Speech-to-text API

Beschrijving

Google Cloud Speech-to-text API is Google's kijk op spraakherkenningssoftware. De API is heel betrouwbaar. Zo heeft het volgens Anggraini e.a. (2018) een succespercentage van 100% voor normale stemmen en een percentage tussen 83,3% en 90% voor mensen met een spraakbeperking. Daarnaast biedt het ook de mogelijkheid om het model uit te breiden door het nieuwe vocabulaire aan te leren.

Functies

De service komt met verschillende functies. Dit zijn de belangrijke kenmerken die ze adverteren:

- **Taalherkenning:** De API ondersteunt spraakherkenning in verschillende talen en dialecten, waardoor gebruikers wereldwijd toegang hebben tot de dienst. Het kan automatisch de gesproken taal detecteren zonder voorafgaande taalconfiguratie.
- **Real-time streaming:** De API ondersteunt real-time streaming van spraak naar tekst, waardoor ontwikkelaars toepassingen kunnen bouwen die directe feedback en interactie vereisen, zoals live ondertiteling, spraakgestuurde commando's en real-time transcriberen.
- **Hoge nauwkeurigheid:** Dankzij de geavanceerde neurale netwerkmodellen levert de Speech-to-Text API nauwkeurige transcripties, zelfs in omgevingen met achtergrondgeluiden, verschillende spreekstijlen en spraak van meerdere sprekers.

- **Aanpasbare modellen:** Gebruikers kunnen de spraakherkenning aanpassen aan specifieke woordenschat en domeinen door aangepaste woordenlijsten en grammatica's te gebruiken. Dit verhoogt de nauwkeurigheid en zorgt voor een betere herkenning van vakspecifieke terminologie.
- **Multimediaondersteuning:** De API kan spraak herkennen in verschillende formaten, waaronder audio-opnamen, audiobestanden en streaming-audio. Dit maakt integratie met verschillende bronnen en toepassingen mogelijk.
- **Beveiliging en privacy:** Google Cloud hanteert strenge beveiligingsmaatregelen om de vertrouwelijkheid en integriteit van de gegevens te waarborgen. De API ondersteunt ook automatische verwijdering van persoonlijk identificeerbare informatie (PII) uit de gegenereerde transcripties.
- **Spraakherkenning op apparaat:** Naast de cloudgebaseerde API biedt Google Cloud ook een on-device spraakherkenning SDK aan, genaamd Speech-to-Text On-Device. Hiermee kunnen ontwikkelaars spraakherkenning rechtstreeks op apparaten uitvoeren, zonder een internetverbinding te vereisen.

Tarieven

De service kan per maand 60 minuten gratis gebruikt worden. Daarna moet er betaald worden per minuut van €0,016 voor standaard met datalogging tot €0,072 voor medisch gebruik zonder datalogging.

3.1.2. Amazon Transcribe

Beschrijving

Amazon Web Services (AWS) biedt een ASR software aan genaamd Amazon Transcribe.

Functies

De functies die Amazon Transcribe aanbied zijn:

- **Taalherkenning:** Amazon Transcribe kan automatisch de gesproken taal detecteren, waardoor het mogelijk is om meertalige spraakherkenning te realiseren zonder voorafgaande taalconfiguratie.
- **Real-time streaming:** Amazon Transcribe heeft real-time streaming van spraak naar tekst, wat ideaal is voor toepassingen zoals live ondertiteling, spraakgestuurde opdrachten en interactieve dialoogsysteem.
- **Overzichtelijke transcripties:** Dit wordt gedaan door het toevoegen van tijds-aanduidingen en het herkennen van meerdere stemmen.

- **Aanpassing van modellen:** Gebruikers kunnen de spraakherkenning verbeteren en aanpassen aan specifieke vocabulaires en domeinen door aangepaste taalmodellen te maken. Hierdoor kunnen ze de nauwkeurigheid van de herkenning verhogen voor gespecialiseerde toepassingen.
- **Privacy:** Amazon Transcribe garandeert beveiligde dataverkeer. Daarnaast is het ook mogelijk woorden uit de transcriptie te filteren als deze ongewenst zijn.

Tarieven

De service kan 60 minuten per maand gratis gebruikt worden, gedurende 12 maanden. Daarna moet er een vergoeding worden betaald afhankelijk van het totale gebruik per maand en de geolocatie van de gebruikte server (Figuur 3.1).

Region: Europe (Paris) ↕		
Tier	Volume (minutes/month)	Standard Batch Transcription (\$/minute)*
T1	First 250,000 minutes	\$0.02400
T2	Next 750,000 minutes	\$0.01500
T3	Next 4,000,000 minutes	\$0.01020
T4	Over 5,000,000 minutes	\$0.00900

Figuur (3.1)

Prijzen Amazon Transcribe voor Parijse server (Amazon, 2023)

3.1.3. Microsoft Azure Speech Services

Beschrijving

Microsoft Azure Speech Services is een spraakherkenningsservice die wordt aangeboden door Microsoft Azure, het cloudplatform van Microsoft.

Functies

Azure Speech Services biedt een breed scala aan functies voor spraakherkenning en tekst-naar-spraak-conversie. Enkele van de belangrijkste functies zijn:

- **Spraakherkenning:** De service kan gesproken taal detecteren en omzetten naar tekst in meer dan 60 talen. Het ondersteunt zowel real-time verwerking als batchverwerking van audiobestanden.
- **Taalherkenning:** Azure Speech Services kan automatisch de gesproken taal detecteren, waardoor het mogelijk is om meertalige spraakherkenning te realiseren zonder voorafgaande taalconfiguratie.
- **Aanpassing van modellen:** Gebruikers kunnen de spraakherkenning verbeteren en aanpassen aan specifieke vocabulaires en domeinen door aangepaste taalmodellen te maken.

paste taalmodellen te maken. Hierdoor kunnen ze de nauwkeurigheid van de herkenning verhogen voor gespecialiseerde toepassingen.

- **Real-time streaming:** Azure Speech Services ondersteunt real-time streaming van spraak naar tekst, wat ideaal is voor toepassingen zoals live ondertiteling, spraakgestuurde opdrachten en interactieve dialoogsystemen.
- **Text-to-speech:** Naast spraakherkenning biedt de service ook mogelijkheden voor tekst-naar-spraak-conversie. Gebruikers kunnen tekst invoeren en de service genereert menselijke spraakuitvoer in verschillende stemmen en talen.

Tarieven

Per maand bied Azure vijf uur gratis aan daarna moet er een vergoeding van €0,935 betaald worden per uur.

3.1.4. Mozilla DeepSpeech

Beschrijving

In tegenstelling tot de voorgaande services is DeepSpeech van Mozilla open source.

Functies

Mozilla DeepSpeech biedt verschillende functies en mogelijkheden voor spraakherkenning. Enkele van de belangrijkste kenmerken zijn:

- **Hoge nauwkeurigheid:** DeepSpeech streeft naar hoge nauwkeurigheid in spraakherkenning en heeft aanzienlijke vooruitgang geboekt in termen van herkenningsscores.
- **Flexibiliteit:** Het systeem kan worden aangepast en afgestemd op specifieke toepassingen of vocabulaires. Hierdoor kan het worden gebruikt voor uiteenlopende spraakherkenningsscenario's.
- **Real-time verwerking:** DeepSpeech is ontworpen om spraak in realtime om te zetten, wat betekent dat het geschikt is voor toepassingen die onmiddellijke spraak-naar-tekstfunctionaliteit vereisen.
- **Privacy:** Als open-sourceproject stelt DeepSpeech gebruikers in staat om spraakherkenningssystemen lokaal uit te voeren, waardoor de privacy van gebruikersinformatie beter gewaarborgd is.

Tarieven

Aangezien dat de service open source is, moet er geen vast tarief betaald worden voor de service. Er kunnen echter wel kosten komen om de service te hosten.

3.1.5. Whisper

Beschrijving

Whisper is de ASR service gemaakt door OpenAI. Het is beschikbaar als een API die ontwikkelaars kunnen integreren in hun eigen toepassingen en systemen door er spraakopnamen naar te sturen. De API geeft vervolgens een tekstuele weergave van de spraak terug als respons.

Functies

Whisper heeft verschillende kenmerken en mogelijkheden die het tot een krachtig spraakherkenningssysteem maken:

- **Hoge nauwkeurigheid:** Whisper heeft aanzienlijke vooruitgang geboekt in termen van spraakherkenningsscores en streeft naar hoge nauwkeurigheid bij het omzetten van spraak naar tekst.
- **Meertaligheid:** Het model van Whisper ondersteunt meerdere talen, waardoor het geschikt is voor wereldwijde toepassingen en gebruikers met verschillende taalvereisten.
- **Flexibiliteit:** Whisper kan worden aangepast en afgestemd op specifieke toepassingen en vocabulaires, waardoor het kan worden gebruikt in uiteenlopende spraakherkenningsscenario's.
- **Real-time verwerking:** Het systeem is ontworpen om spraak in realtime te verwerken, wat betekent dat het geschikt is voor toepassingen die onmiddellijke spraak-naar-tekstfunctionaliteit vereisen.
- **Privacy:** OpenAI hanteert strenge beveiligings- en privacyrichtlijnen om de gebruikersgegevens te beschermen. Whisper stelt gebruikers in staat om spraakherkenningssystemen lokaal uit te voeren, waardoor de privacy van gebruikersinformatie beter gewaarborgd is.

Tarieven

Net als Mozilla's DeepSpeech is Whisper open source. Dit betekent dat de kosten afhangen van waar en hoe de service wordt gehost.

3.1.6. Kaldi ASR

Beschrijving

Kaldi ASR is een open-source spraakherkenningssysteem dat algemeen wordt gebruikt in de academische wereld en de industrie. Het is ontwikkeld door het Kaldi-project, een community-gedreven initiatief dat zich richt op het leveren van state-of-the-art spraaktechnologie.

Kaldi maakt gebruik van geavanceerde algoritmen en technieken, waaronder deep neural networks (DNN's) en hidden Markov models (HMM's), om spraak naar tekst

om te zetten. Het biedt een flexibel en configureerbaar framework dat kan worden aangepast aan verschillende spraakherkenningstoepassingen.

Functies

Kaldi ASR biedt een breed scala aan functies en mogelijkheden die het tot een krachtig spraakherkenningssysteem maken:

- **Hoge nauwkeurigheid:** Kaldi is bekend om zijn hoge nauwkeurigheid in spraakherkenning. Het maakt gebruik van geavanceerde modellen en trainingstechnieken om nauwkeurige resultaten te leveren.
- **Flexibiliteit:** Het Kaldi-framework is zeer configureerbaar en aanpasbaar. Het stelt gebruikers in staat om modellen en akoestische kenmerken aan te passen aan specifieke toepassingen en datasets.
- **Meertaligheid:** Kaldi ondersteunt meerdere talen en kan worden gebruikt voor spraakherkenning in verschillende taalomgevingen.
- **Uitgebreide toolkit:** Kaldi wordt geleverd met een uitgebreide toolkit die verschillende hulpmiddelen en utilities biedt voor spraakdataverwerking, modeltraining en evaluatie.
- **Community-ondersteuning:** Het Kaldi-project wordt ondersteund door een actieve gemeenschap van ontwikkelaars en onderzoekers, wat resulteert in regelmatige updates, bugfixes en nieuwe functies.

Gebruik en implementatie

Kaldi ASR wordt meestal gebruikt via de command line-interface en vereist enige technische kennis om effectief te kunnen gebruiken. Het proces om Kaldi te gebruiken omvat het verzamelen en voorbereiden van spraakgegevens, het trainen van akoestische en taalmodellen, en het uitvoeren van de spraakherkenning op nieuwe gegevens.

Kaldi biedt documentatie en handleidingen om gebruikers te begeleiden bij de implementatie en het gebruik van het systeem. Het heeft een actieve gebruikersgemeenschap waar gebruikers vragen kunnen stellen en ervaringen kunnen delen.

Beschikbaarheid en prijs

Kaldi ASR is een open-source project en is vrij beschikbaar voor iedereen. Het kan worden gedownload van de officiële Kaldi-website en lokaal worden uitgevoerd.

3.2. Natuurlijke taalverwerking

3.2.1. GPT series

Beschrijving

The GPT, wat staat voor Generative Pre-trained Transformer, series zijn algemene taalmodellen gemaakt door OpenAI. Het doel van deze modellen is genereren van menselijke taal op basis van de gegeven input. Het meest bekende voorbeeld is OpenAI's chatbot, ChatGPT. Het maakt gebruik van GPT-3.5 en GPT-4 om antwoorden te genereren. Deze modellen zijn getraind op een brede set van internet data.

Tarieven

Er zijn twee GPT-4 modellen beschikbaar de 8K context en de 32K context. Om ze te gebruiken moet er betaald worden per 1000 tokens. Dat is ongeveer gelijk aan 750 woorden. Voor de 8K context is de prijs \$0.03 voor invoer en \$0.06 voor de uitvoer. De kost voor 32K context is \$0.06 voor invoer en \$0.12 voor het resultaat. Afhankelijk van het model is de prijs voor GPT-3 tussen de \$0.0004 of \$0.02 per 1000 tokens.

3.2.2. BERT

Beschrijving

BERT, wat staat voor Bidirectional Encoder Representations from Transformers, is een ander type taalmodel ontwikkeld door Google AI Language. In tegenstelling tot de GPT-series, is BERT een bi-richtingstransformer, wat betekent dat het in staat is om zowel de context vóór als na een woord te begrijpen tijdens het trainen.

Tarieven

In tegenstelling tot de GPT reeks is BERT gratis te gebruiken. Het taalmodel vereist wel aanzienlijke rekenkracht en geheugen waardoor de hardware waarop het draait sterk genoeg moet zijn.

3.2.3. FlairNLP

Beschrijving

FlairNLP is een NLP (Natural Language Processing) framework ontwikkeld door Zalando Research. Het staat voor "Flexible Language-Agnostic IRst-orderen is gericht op het bieden van flexibiliteit en veelzijdigheid in het verwerken van natuurlijke taal. Net als BERT en GPT maakt FlairNLP gebruik van transformer-gebaseerde modellen en kan worden gebruikt voor een breed scala aan taalverwerkings-taken.

Tarieven

Ook FlairNLP is open-source wat wil zeggen dat de kosten afhangen van hoe en waar de software op draait.

3.2.4. RoBERTa**Beschrijving**

RoBERTa staat voor "A Robustly Optimized BERT Pretraining Approach." Het is een door Meta AI ontwikkelde transformer-gebaseerd taalmodel en een doorontwikkeling van BERT. RoBERTa is ontworpen om de prestaties van BERT verder te verbeteren door middel van optimalisaties in de trainingsprocedure.

In tegenstelling tot BERT maakt RoBERTa gebruik van een grotere hoeveelheid trainingsdata en een langere trainingsduur, wat het model helpt om een dieper taalbegrip te ontwikkelen. Het trainingsproces van RoBERTa omvat het maskeren van woorden in een zin en het uitdagen van het model om de ontbrekende woorden correct te voorspellen, vergelijkbaar met BERT.

Tarieven

RoBERTa is net als BERT gratis te gebruiken.

4

Methodologie

Dit onderzoek start met een literatuurstudie (2). Vervolgens wordt er een requirementsanalyse gehouden waar gekeken wordt wat er van de oplossing verwacht wordt (4.1). Na de requirements vast te leggen zullen verschillende spraak naar tekst software (??) en taalverwerkingsmodellen (??) getest worden. Als laatste zal er een proof-of-concept opgesteld worden.

4.1. Requirementsanalyse

De requirementsanalyse start met een opsomming van functionele en niet-functionele requirements. Vervolgens worden ze aan de hand van de MoSCoW methode gerangschikt op relevantie. Als laatste wordt er een use case opgesteld.

4.1.1. Functionele requirements

- De applicatie moet in staat zijn om spraak te herkennen en te registreren hoe lang en tot wanneer de gebruiker spreekt.
- De applicatie moet in staat zijn om de gesproken woorden om te zetten in tekst, zodat deze kunnen worden opgeslagen en verwerkt.
- De applicatie moet in staat zijn om op basis van de gegenereerde tekst een volgend fragment te kiezen dat aansluit bij het onderwerp en de context van de tekst.
- De applicatie moet gebruik maken van AI-technieken om spraakherkenning en tekstgeneratie mogelijk te maken.
- De applicatie moet gemakkelijk te gebruiken zijn voor de gebruiker, zonder dat deze technische kennis nodig heeft.

- De applicatie moet betrouwbaar zijn en weinig tot geen fouten maken bij het herkennen van spraak en genereren van tekst.
- De applicatie moet in staat zijn om de opgeslagen tekst te doorzoeken op bepaalde trefwoorden of zinnen, om zo snel specifieke informatie te kunnen vinden.
- De applicatie moet aanpasbaar zijn aan de specifieke wensen en behoeften van de gebruiker, zoals het leren herkennen van patronen in gebruikers met een spraakbeperking.

4.1.2. Niet-functionele requirements

NFR	Betrouwbaarheid
Indicator	Volwassenheid
Meetvoorschrift	De spraakherkenningssoftware moet een nauwkeurighedsniveau van minimaal 95% bereiken bij het herkennen en transcriberen van spraak binnen de eerste 6 maanden na de lancering.
Norm	De spraakherkenningssoftware moet zeer nauwkeurig zijn bij het herkennen en transcriberen van spraak, omdat eventuele fouten kunnen leiden tot onnauwkeurige resultaten en mogelijk onjuiste interpretaties.
NFR	Prestatie-efficiëntie
Indicator	Snelheid
Meetvoorschrift	De spraakherkenningssoftware moet spraak in real-time verwerken met een maximale vertraging van 1 seconde tussen het uitspreken van de zin en het verschijnen van de transcriptie bij de lancering en deze snelheid behouden.
Norm	De software moet snel genoeg zijn om de spraak te herkennen en te transcriberen, zodat het in real-time kan werken en niet leidt tot onnodige vertragingen of onderbrekingen.
NFR	Beveiligbaarheid
Indicator	Vertrouwelijkheid
Meetvoorschrift	De spraakgegevens van gebruikers moeten worden versleuteld tijdens overdracht en opslag, en alle dataoverdracht en opslag moeten bij de lancering beveiligd zijn volgens de best practices, met regelmatige beveiligingsaudits.
Norm	De privacy en veiligheid van de gebruiker moeten worden beschermd, zodat de spraakgegevens niet kunnen worden gehackt of gelekt.

4.1.3. MoSCoW

Aan de hand van de MoSCoW methode bepalen we het belang van verschillende functionaliteiten.

Must have	De applicatie moet in staat zijn om de gesproken woorden om te zetten in tekst, zodat deze kunnen worden opgeslagen en verwerkt.
	De applicatie moet in staat zijn om op basis van de gegenereerde tekst een volgend fragment te kiezen dat aansluit bij het onderwerp en de context van de tekst.
	De applicatie moet gebruik maken van AI-technieken om spraakherkenning en tekstgeneratie mogelijk te maken.
	De applicatie moet in staat zijn om de opgeslagen tekst te doorzoeken op bepaalde trefwoorden of zinnen, om zo snel specifieke informatie te kunnen vinden.
Should have	De applicatie moet in staat zijn om spraak te herkennen en te registreren hoelang en tot wanneer de gebruiker spreekt.
	De applicatie moet gemakkelijk te gebruiken zijn voor de gebruiker, zonder dat deze technische kennis nodig heeft.
	De applicatie moet betrouwbaar zijn en weinig tot geen fouten maken bij het herkennen van spraak en genereren van tekst.
Could have	De applicatie moet aanpasbaar zijn aan de specifieke wensen en behoeften van de gebruiker, zoals het leren herkennen van patronen in gebruikers met een spraakbeperking.
	Een flexibel systeem met verschillende scenario's creëren.

4.1.4. Use Case

4.2. Proof-of-concept

dfsdfsd

5

Conclusie



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

In de reclamespot 'The Impact Will Be Real' over de Metaverse toont Meta ([2022](#)) hun visie over de rol die Virtual Reality (VR) speelt in de toekomst. Zo laten ze verschillende toepassingen ervan in het onderwijs zien. Jammer genoeg staat onze technologie nog niet zo ver als wat er in de reclame gezien kan worden maar ook nu al vindt VR zich een baan in verschillende opleidingen.

De Hogeschool van Gent maakt ook gebruik van VR om studenten de kans te geven in meer realistische situaties te oefenen. Hiervoor zijn twee verschillende technieken gebruikt. Allereerst heb je het renderen van een omgeving. Dit laat de gebruiker een interactieve wereld van 3D modellen ontdekken. Zo bestaan er drie virtuele kamers waarin de student kan oefenen. De andere manier is aan de hand van een 360° opname die wordt gemaakt aan de hand van een 360° camera. Omdat dit een opname van de werkelijkheid neemt, ziet deze methode er realistischer uit.

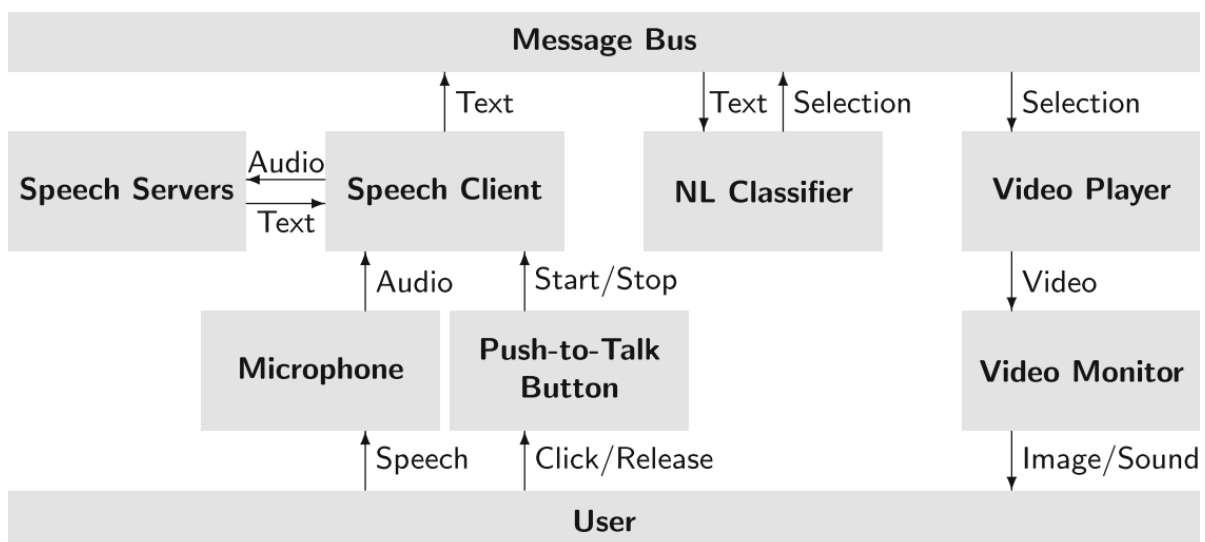
Dit is waar er op het probleem wordt gestoten. Aangezien de tweede methode werkt met een opname moet er naar verschillende fragmenten gesprongen worden naargelang het antwoord dat de gebruiker ingeeft. Dit wordt handmatig gedaan via een meerkeuzevraag. Hierdoor staat de oefening tijdelijk stil wat de echtheid van de situatie weghaalt. Daarom wordt in deze bachelorproef toegepaste informatica onderzocht hoe het overschakelen anders kan aangepakt worden zodat het voor de gebruiker realistischer aanvoelt. Hiervoor kijken we richting Artificiële Intelligentie (AI).

Alle grote IT bedrijven hebben wel een departement die zich bezighoudt met AI. Ze zien allemaal de mogelijkheden die het biedt. Van jobs gemakkelijker te maken tot het zelf creëren van kunst, de toepassingen zijn oneindig. Daarom zal in functie van het Zorglab op zoek worden gegaan naar een Spraak naar tekst (STT) en een natuurlijke taalverwerking software (NLP).

De STT zal worden gebruikt om het manueel ingeven van het antwoord te vervangen. In plaats daarvan zal de gebruiker gewoon zijn antwoord luidop kunnen geven en zal de STT dit in tekst omzetten. Dit alleen zal natuurlijk geen volgend fragment voor de gebruiker kunnen kiezen en daarom hebben we de NLP nodig. Deze zal de gegenereerde tekst omzetten naar kernwoorden en aan de hand daarvan bepalen welk fragment als volgende geschikt is.

A.2. State-of-the-art

Om de immersie in de simulatie te vergroten moet ervoor gezorgd worden dat het overschakelen naar een volgend fragment kan gebeuren zonder de gebruiker uit de illusie te halen. Hiervoor zouden verbale antwoorden gegeven kunnen worden. Een mooi voorbeeld van zo een interactie is het project 'Dimensions in Testimony' van de USC Shoah Foundation (2020). Daar kunnen bezoekers vragen stellen aan een overlevende van de Holocaust die op voorhand een volledig interview heeft afgelegd. Dit werd mogelijk gemaakt dankzij het systeem dat erachter zit (Figuur A.1). Dit systeem is opgebouwd uit volgende componenten: een software voor spraakherkenning (ASR) die de gebruikers verbale vraag in tekst omzet, een Natural Language Classifier (NLC) die op basis van de gegenereerde tekst een antwoord via een audio/video fragment voorziet en een mediaspeler die de fragmenten kan afspelen met tussendoor een inactieve animatie (Traum e.a., 2015).



Figuur (A.1)

Dimensions in Testimony - Systeem Architectuur (Traum e.a., 2015)

Spraakherkenning is de laatste jaren veel vooruitgegaan, dit komt omdat er meer data beschikbaar is, de algoritmes beter (diep neural networks) en de computers sneller zijn. Maar perfect zullen ze nooit zijn zegt van Hessen (2020). Dit is geen verrassing aangezien mensen zelf vaak nog problemen hebben bij het verstaan van een andere. De meest voorkomende fouten van een ASR zijn herkenningsfouten. Deze kunnen ontstaan door slechte kwaliteit van de opname maar ook van de manier waarop woorden worden uitgesproken. De andere fouten ontstaan omdat ASR zijn vocabulaire niet uitgebreid genoeg is. Dit noemen we Out-of-vocabulary-fouten (OOV) (van Hessen, 2020).

Naast de ASR hebben we ook de natuurlijke taalverwerking (NLP). In 'Dimension in Testimony' maakten ze hiervoor gebruik van NCPeditor (Figuur A.2). De classifier in dit systeem berekent welke antwoorden op de ingegeven tekst kunnen gegeven worden door de taalmodellen van beide te vergelijken en de antwoorden te rangschikken. Zolang de ingegeven tekst een foutmarge lager dan 50% blijft, zal het antwoord ongewijzigd blijven (Leuski & Traum, 2010).

A.3. Methodologie

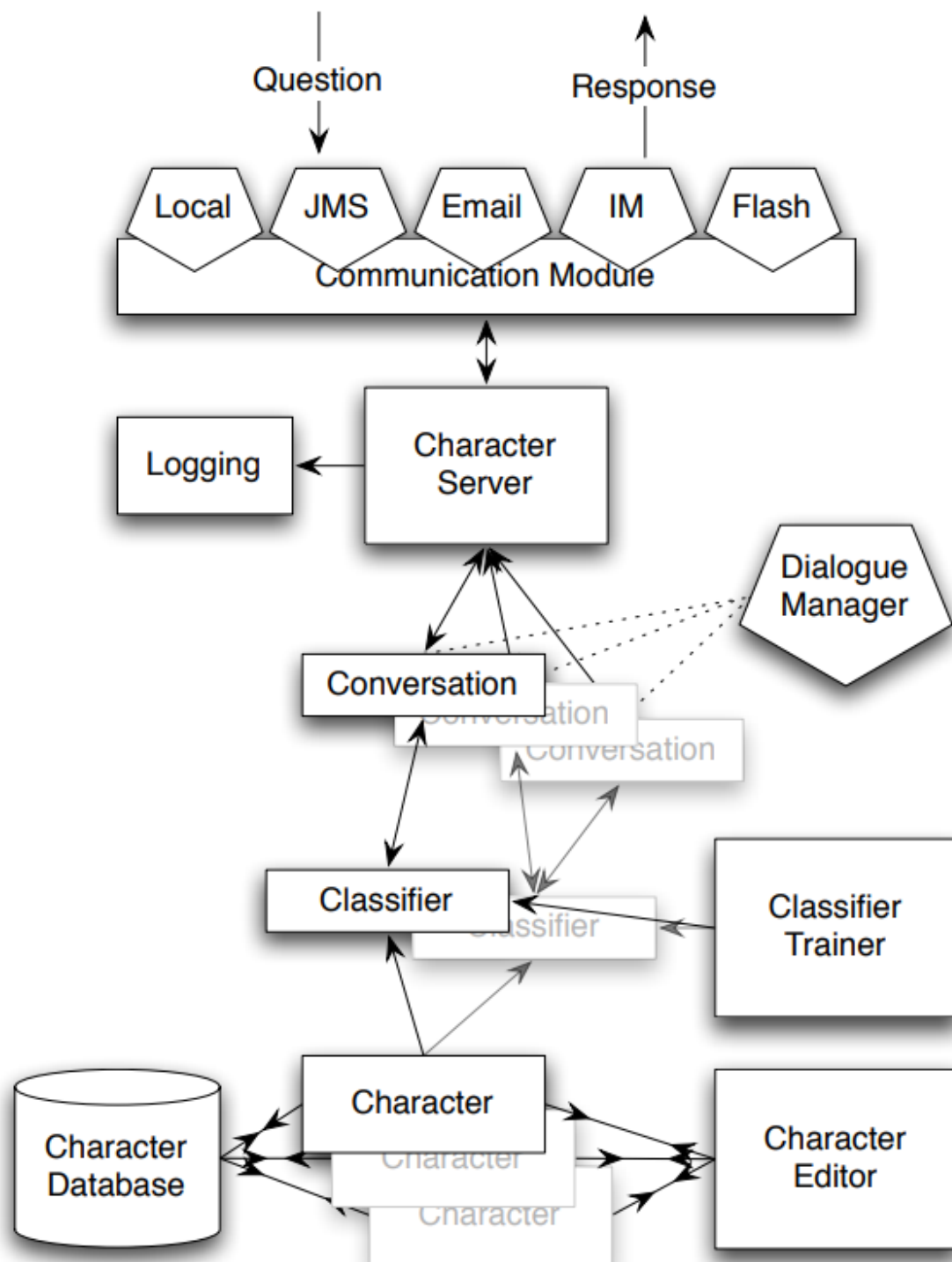
Om te beginnen zal er een uitgebreide literatuurstudie worden gehouden. Deze zal een overzicht bevatten van de vooruitgang en huidige stand van VR, ASR en NLP. Naar het einde van de literatuurstudie toe zal ook een long list worden opgesteld. Deze lijst zal bestaan uit bestaande ASR en NLP software. Voor elk item zal minimum een beschrijving, hun functies en tarieven bevatten. Ook zal er worden samengezeten met het team van het Zorglab om een requirementsanalyse uit te voeren. Zo kan worden neergeschreven wat de problemen zijn en waar ze tegenaan lopen. Daarin wordt besproken aan de hand van de MoSCoW methode wat er verwacht wordt van een oplossing, hoe moet deze in zijn werk gaan? Op basis daarvan zal de long list ingekort worden tot een short list worden opgesteld. Er zal bijvoorbeeld gekeken worden welke ASR's er al verkrijgbaar zijn en welke er in het Nederlands werken.

Wanneer er een uitgebreide literatuurstudie is uitgevoerd, zal een proof of concept opgesteld worden. Hierin wordt de software die gevonden is op de proef gesteld. Hier wordt gekeken hoe correct de ASR's inkomende audio kunnen transcriberen bij personen met een stotter. Daarnaast zal ook getest worden of er aan de hand van kernwoorden een juist fragment gekozen kan worden. Door de AI verschillende prompts te geven en te kijken of hij naar de juiste tijdsaanduiding in de video gaat. Nadat de poc afgewerkt is, wordt deze beoordeelt om te kijken hoe het in de context van het Zorglab toegepast kan worden.

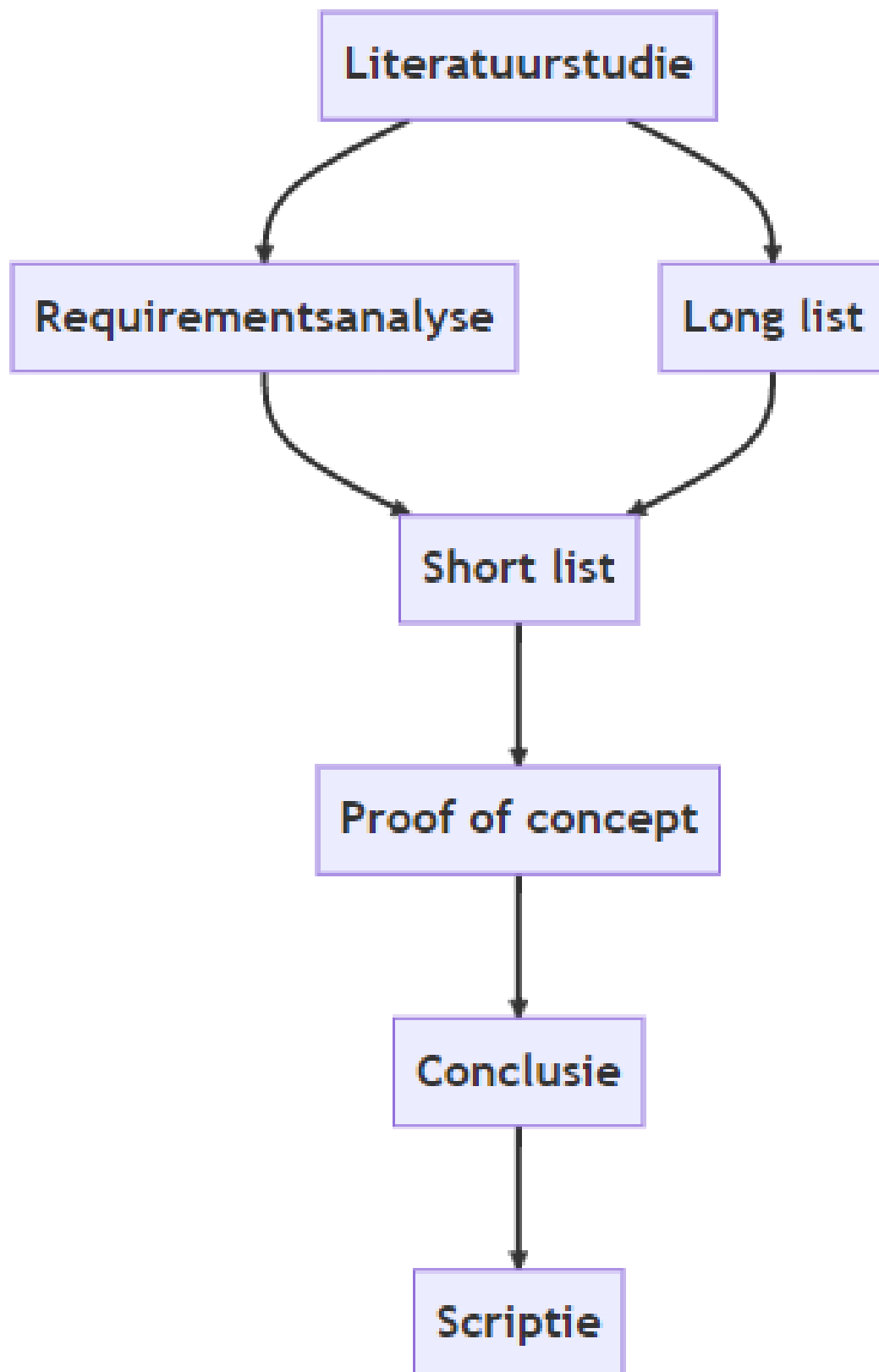
Als laatste zal deze scriptie worden afgewerkt door het schrijven van een voorwoord, de samenvatting en een conclusie. Wanneer dit is voltooid zal deze proef worden beoordeeld waarna een positief resultaat een mondelinge verdediging zal plaatsvinden.

A.4. Verwacht resultaat, conclusie

Uit deze bachelorproef wordt verwacht dat er een proof of concept opgesteld wordt. Hierin zal gekeken zijn of het haalbaar is om aan de hand van spraakherkenning en natuurlijke taalverwerking software stottertherapie oefening, in het Zorglab, vlotter te doen verlopen. Zo kan het Zorglab kiezen of ze dit idee effectief aan het lab willen toevoegen. Dit kan dan aan de hand van de software dat in het PoC is gebruikt. Anderzijds kunnen ze op basis ervan andere software met hetzelfde doeleinde inzetten.

**Figuur (A.2)**

NPCEditor - Systeem Architectuur (Leuski & Traum, 2010)



Figuur (A.3)
Flowchart methodologie

Bibliografie

- Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, 9, 131858–131876. <https://doi.org/10.1109/ACCESS.2021.3112535>
- Amazon. (2023, januari 1). *Amazon Transcribe: Developer Guide*. <https://aws.amazon.com/transcribe>
- Anggraini, N., Kuniawan, A., Wardhani, L., & Hakiem, N. (2018). Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API. *Telkomnika (Telecommunication Computing Electronics and Control)*, 16, 2733–2739. <https://doi.org/10.12928/TELKOMNIKA.v16i6.9638>
- Boas, Y. A. G. V. (2012). Overview of Virtual Reality Technologies.
- Bryson, S. (2013). Virtual Reality: A Definition History - A Personal Essay.
- Chee, L. S., Ai, O. C., Yaacob, S. B., & Perlis, J. (2009). Overview of Automatic Stuttering Recognition System.
- Gordon, N. (2002). Stuttering: incidence and causes. *Developmental Medicine and Child Neurology*, 44(4), 278–282. <https://doi.org/10.1017/S0012162201002067>
- Leuski, A., & Traum, D. (2010). NPCEditor: A Tool for Building Question-Answering Characters. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Verkregen december 15, 2022, van http://www.lrec-conf.org/proceedings/lrec2010/pdf/660_Paper.pdf
- Meta. (2022, juni 15). *The Impact Will Be Real*. Verkregen december 11, 2022, van <https://www.youtube.com/watch?v=80IIEnSNwQc>
- Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., Debevec, P., Gainer, A., Georgila, K., Haase, K., Jungblut, K., Leuski, A., Smith, S., & Swartout, W. (2015). New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling. In H. Schoenau-Fog, L. E. Bruni, S. Louchart & S. Baceviciute (Red.), *Interactive Storytelling* (pp. 269–281). Springer International Publishing.
- USC Shoah Foundation. (2020, april 3). *Dimensions in Testimony | USC Shoah Foundation*. Verkregen december 5, 2022, van <https://www.youtube.com/watch?v=nGzAc9mIoTM>
- van Hessen, A. (2020, augustus 1). *Automatische spraakherkenning - Hoe kun je het inzetten voor onderwijsmateriaal?* (Onderzoeksrap.). SURF.