Hoe kan het herkennen van verbale antwoorden door spraakherkenning software, de VR ervaring van stotterthe-rapie oefening realistischer doen aanvoelen.

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
Eerste examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Samenvatting

Inhoudsopgave

Lijst van figuren

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 patiënt 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 het moet bijwonen. 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. -Kan de spraakherkenningssoftware registreren hoelang en tot wanneer de gebruiker spreekt. -Kan de spraakherkenningssoftware transcriberen wat er werd gezegd. -Kan de applicatie een volgend fragment kiezen 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 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.
In Hoofdstuk 4, 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. Virtuele realiteit

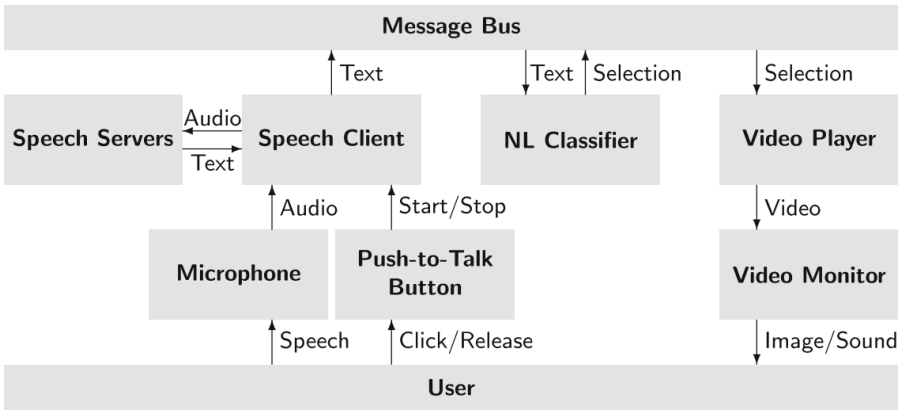
Hoewel Virtual Reality tegenwoordig veel gevorderd is, bestaat het al voor een lange tijd. Het eerste toestel dat de werkelijkheid nabootste was Morton Heilig's Sensorama uit 1962. Dit liet de gebruiker ervaren hoe het voelde om op een motor door Boston te rijden. De opvolgende jaren werden ook andere 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 een 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.

2.2. 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 2.2.2)
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.

2.2.1. Spraakherkening

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. 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).

Google Cloud Speech-to-text API

Google Cloud Speech-to-text API is Google's kijk op spraakherkennigssoftware. Hun service kan per maand 60 minuten gratis gebruikt worden. Daarna moet er per minuut van €0,016 voor standaard met datalogging tot €0,072 voor medisch gebruik zonder datalogging. De service komt ook met verschillende functies zoals het automatisch detecteren van de gesproken taal, het herkennen van verschillende stemmen en real-time streaming om live spraak naar tekst om te zetten.
De API is ook heel betrouwbaar. Zo heeft een succespercentage van 100% voor normale stemmen en een percentage tussen 83,3% en 90% voor mensen met een spraakbeperking (Anggraini e.a., 2018). Daarnaast bied het ook de mogelijkheid om het model uit te breiden door het nieuwe vocabulaire aan te leren.

Amazon Transcribe

test

Microsoft Azure Speech Services

test dfsdf

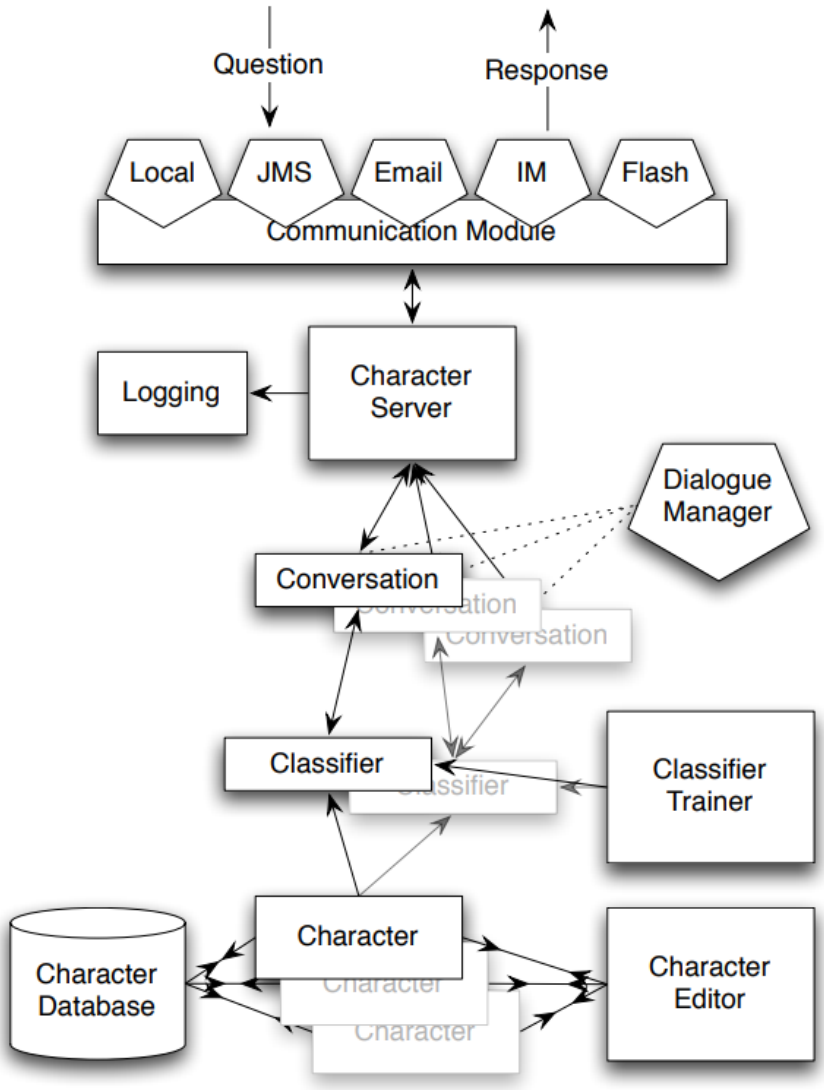
Mozilla Deepspeech

test

Whisper
test

2.2.2. Natuurlijke taalverwerking

Naast de ASR hebben we ook de natuurlijke taalverwerking (NLP). 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

Methodologie

Dit onderzoek start met een literatuurstudie (2). Vervolgens wordt er een requirementsanalyse gehouden waar gekeken wordt wat er van de oplossing verwacht wordt (3.1). Na de requirements vast te leggen zullen verschillende spraak naar tekst software (3.2) en taalverwerkingsmodellen (3.3) getest worden. Als laatste zal er een proof-of-concept opgesteld worden.

3.1. Requirementsanalyse

dfdsf

3.1.1. Use Cases

3.1.2. MoSCoW

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

1. Must have:
    - Het detecteren van wanneer de gebruiker spreekt.
    - Het transcriberen van wat er gezegd wordt.
2. Should have:
    - Aan de hand van de gegenereerde tekst het volgend fragment bepalen.
    - The text in the entries may be of any length
3. Could have:
    - Een flexibel systeem met verschillende scenario's creëren

3.2. Spraak naar tekst

dfsdfsdf

3.3. Taalmodellen

dfsdfsd

3.4. Proof-of-concept

dfsdfsd

4

Conclusie

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.

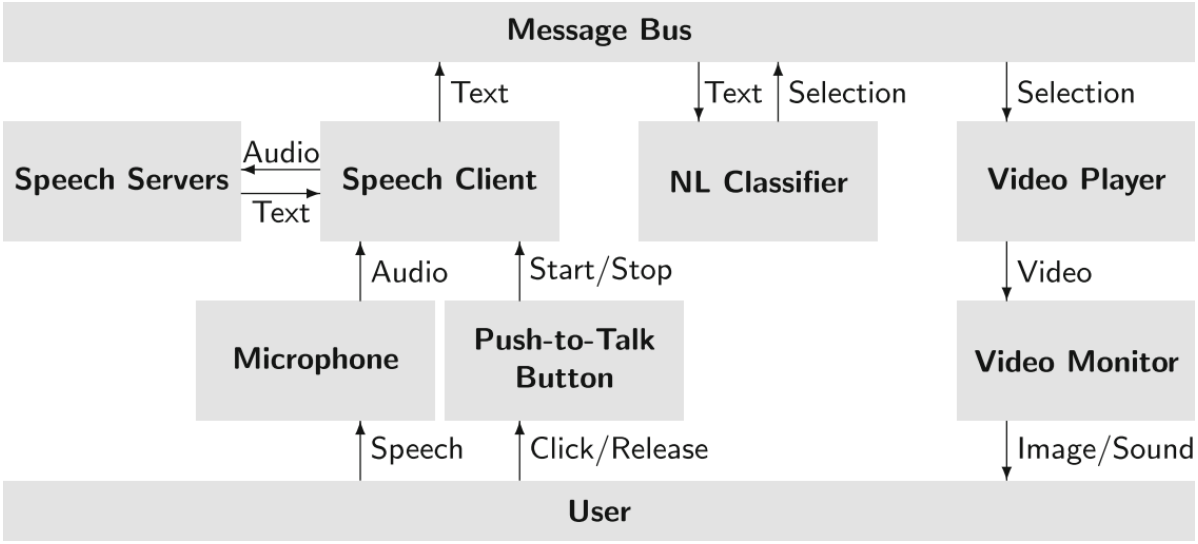
A.1. Introductie

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 (deep 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).

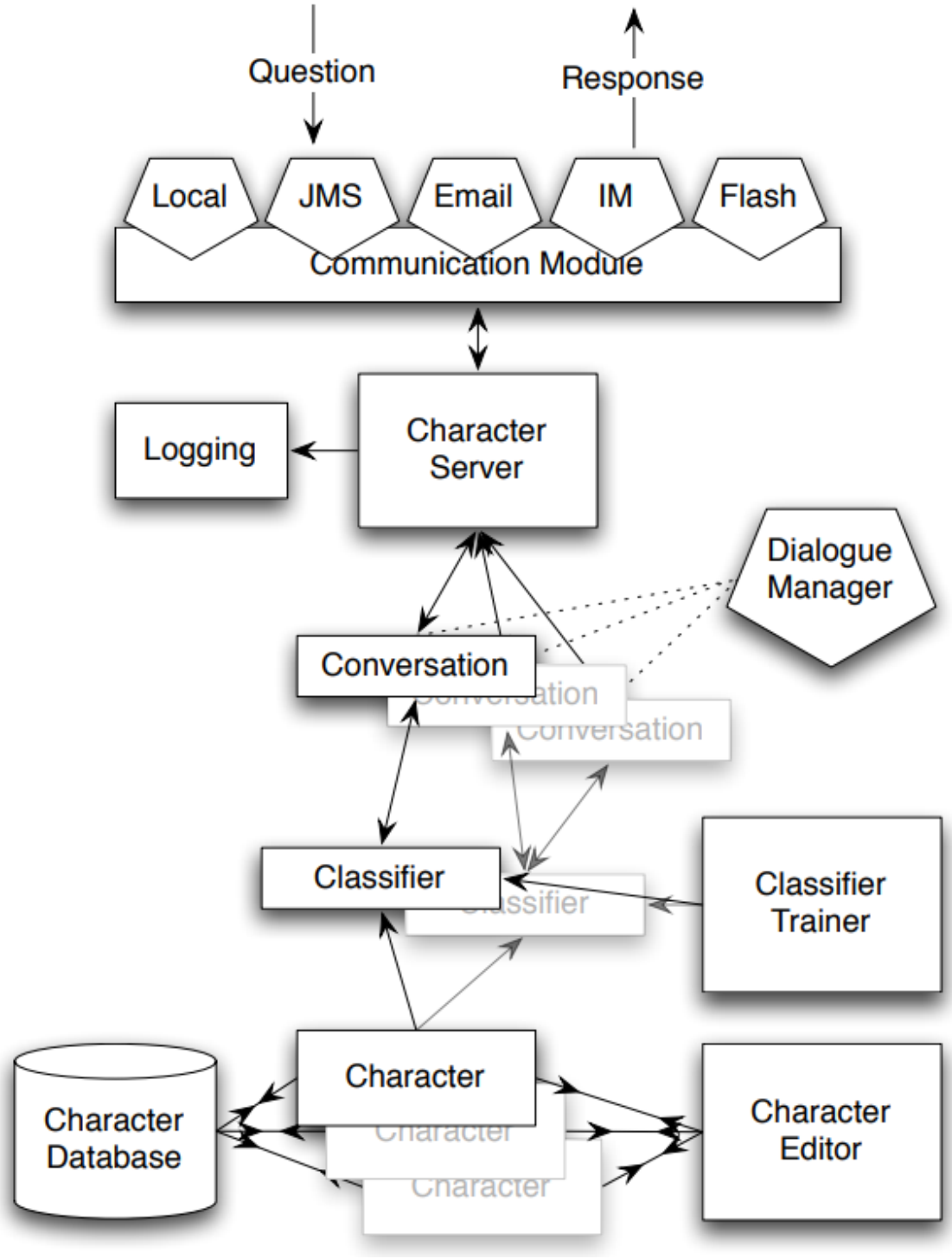
A.3. Methodologie

Om te beginnen zal er 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. Ook wordt daarin 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 literatuurstudie uitgebreid worden met onderzoek naar oplossingen die dit probleem kunnen verhelpen. Hieronder valt onder andere een onderzoek naar ASR. Er zal gekeken worden welke ASR's er al verkrijgbaar zijn en welke er in het Nederlands werken. Alleen een ASR hebben is niet genoeg om de immersie te behouden dus zal er ook worden gekeken naar hoe, aan de hand van de gegenereerde tekst, het volgende fragment getoond kan worden.
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.

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)

Bibliografie

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.

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.