TIETO- JA SÄHKÖTEKNIIKAN TIEDEKUNTA

Ida Haataja
Veikko Romppainen
Juha-Matti Runtti

HUMANOIDIROBOTIN PUHEENTUNNISTUSSOVELLUS MELUISASSA YMPÄRISTÖSSÄ

Kandidaatintyö
Tietotekniikan tutkinto-ohjelma
Toukokuu 2024

Haataja I., Romppainen V., Runtti J. (2024) Humanoidirobotin puheentunnistussovellus meluisassa ympäristössä. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 33 s.

TIIVISTELMÄ

Ihmisten ja robottien välinen vuorovaikutus sekä sosiaaliset robotit ovat kehittyneet merkittävästi vuosien varrella. Robotit kykenevät ymmärtämään puhetta sekä keskustelemaan sujuvasti ihmisten kanssa, ja ne toimivatkin jo laajasti eri tehtävien, kuten asiakaspalvelun parissa. Tärkeä osa sosiaalisia robotteja ovat toimiva puheentunnistusjärjestelmä kuin myös kehittynyt keskustelutekoäly.

Tässä opinnäytetyössä kehitettiin 3d-tulostettuun InMoov-robottiin puheentunnistusjärjestelmä sekä sen kanssa toimiva keskustelutekoäly. Robotin on tarkoitus olla osana esittelemässä tietotekniikan alaa Oulun yliopiston Hakijan päivillä messumaisessa ympäristössä. Ympäristön takia melunsuodatus on tärkeä osa puheentunnistusjärjestelmän toteutusta. Työssä hyödynnetään puheentunnistukseen valmista Python-ohjelmointikielen SpeechRecognition -kirjastoa sekä keskustelutekoälyä varten ChatterBot -kirjastoa.

Työn tuloksena robotille saatiin kehitettyä toimiva puheentunnistusjärjestelmä sekä yksinkertainen sääntöpohjainen keskustelutekoäly, joiden ansiosta robotti kykenee toimimaan sille tarkoitetussa ympäristössä. Järjestelmät tosin vaativat vielä jatkokehitystä melunsuodatuksen ja erityisesti keskustelutekoälyn osalta, jotta robotti hallitsisi laajemman sanavaraston ja siten kykenisi sujuvampaan vuorovaikutukseen ihmisten kanssa. Lisäksi ihmismäisemmän kokemuksen luomiseksi muun muassa robotin eleiden olisi hyvä toimia synkronoidusti puheen kanssa.

Avainsanat: puheentunnistus, keskustelutekoäly, sosiaalinen robotti, melunsuodatus

Haataja I., Romppainen V., Runtti J. (2024) Speech Recognition Application for a Humanoid Robot in a Noisy Environment. University of Oulu, Degree Programme in Computer Science and Engineering, 33 p.

ABSTRACT

The interaction between humans and robots has evolved significantly over the years. Social robots can understand speech and discuss fluently with humans, while working within different tasks in various fields, such as customer service. An essential part of these social robots is a functional speech recognition system, along with well-developed conversational artificial intelligence.

In this thesis, a speech recognition system was developed for a 3d-printed InMoov robot, along with a conversational artificial intelligence system. The main task of the robot is to present information on studying computer science and engineering at the University of Oulu during Applicant Day in a fair-like environment. Due to the noisy environment, an essential part of the system is a functioning noise filtering. For the implementation of this work, the SpeechRecognition library in the Python programming language is used for speech recognition and the ChatterBot library for the conversational artificial intelligence.

As a result of the work, a functional speech recognition system and a simple rule-based conversational artificial intelligence were developed for the robot, allowing it to operate in its intended environment. However, the systems still require further development, especially in noise filtering and conversational artificial intelligence, to enable the robot to better handle broader vocabularies and thus engage in more natural conversations with humans. Also to embrace robot's human-likeness, it would be beneficial for the robot's gestures to synchronize with its speech.

Keywords: chatbot, speech recognition, social robots, noise filtering

SISÄLLYSLUETTELO

TIIVISTELMÄ
ABSTRACT
SISÄLLYSLUETTELO
ALKULAUSE
LYHENTEIDEN JA MERKKIEN SELITYKSET
1. JOHDANTO 7
2. PUHEENTUNNISTUSJÄRJESTELMÄT 8
2.1. Modernit puheentunnistussovellukset 8
2.2. Sosiaaliset robotit 8
2.3. Puheentunnistusjärjestelmän tekninen toteutus 9
2.3.1. Signaalinkäsittely 10
2.3.2. Akustiset mallit 10
2.3.3. Kielimallit 12
2.4. Melunsuodatus 12
2.4.1. Mikrofonin vaikutus melunsuodatukseen 12
2.4.2. Melunsuodatus ohjelmallisesti 14
2.5. Sovellusympäristö 14
3. KESKUSTELUTEKOÄLYT 15
3.1. Ihmisen ja robotin välinen keskustelu 15
3.2. Modernit keskustelutekoälyt 15
3.3. Vastauksen valinta 16
3.4. Keskustelutekoälyjen vaarat 17
4. TOTEUTUS 18
4.1. InMoov-robotti 18
4.2. Puheentunnistusjärjestelmän ominaisuuksien valinta 18
4.3. Puheentunnistusjärjestelmä 19
4.4. Keskustelutekoäly 20
4.4.1. Botin kouluttaminen 21
4.4.2. Esikäsittely 22
4.5. Integraatio 22
4.6. Testaus 23
5. POHDINTA 25
6. JATKOKEHITYS 27
7. PROJEKTIN KUVAUS 28
8. YHTEENVETO 29
9. VIITTEET 30

ALKULAUSE

Työ tehtiin osana Oulun yliopiston Sulautettujen ohjelmistojen projekti -kurssia.

Oulussa 26. toukokuuta 2024

Ida Haataja
Veikko Romppainen
Juha-Matti Runtti

LYHENTEIDEN JA MERKKIEN SELITYKSET

HMM	Hidden Markov Model, Markovin piilomalli
DNNs	Deep Neural Networks, syvät neuroverkot
ROS2	Robot Operating System 2
FFT	Fast Fourier Transform, nopea Fourier-muunnos
STFT	Short Time Fourier Transform
GMM	Gaussian Mixture Model, Gaussin sekoitemalli
DSR	Distant Speech Recognition, kaukainen puheentunnistus
VAD	Voice Activity Detection, puheen aktiivisuuden tunnistus
NLU	Natural-Language Understanding, luonnollisen kielen ymmärtäminen
LLM	Large Language Model, suuri kielimalli
AIML	Artificial Intelligence Markup Language
API	Application Programming Interface, ohjelmointirajapinta
MSMARCO	Microsoft Machine Reading Comprehension

1. JOHDANTO

Robotit ovat olleet osa ihmisten elämää vuosikymmenien ajan. Aikaisemmista yksinkertaisista mekaanisista koneista aina nykypäivän humanoidirobotteihin asti robotteja on käytetty automatisoimaan ja helpottamaan ihmisten jokapäiväistä elämää. Robottien teknologiassa on tehty suuria läpimurtoja viime aikoina - humanoidirobottien puheentunnistus ja puhesynteesi, kasvojentunnistus sekä ihmismäiset ilmeet ja eleet ovat kehittyneet suurta vauhtia. Toimiva puheentunnistus on merkittävässä roolissa ihmisen ja robotin välisessä kommunikaatiossa. Puhesignaalin muuttaminen tietokoneella sanoiksi ja näistä sanoista tarkoituksen ymmärtäminen on ollut tärkeimpiä tutkimuskohteita ihmisen ja tietokoneen välisessä vuorovaikutuksessa [1]. Puheentunnistusteknologiat ovat kehittyneet ensimmäisistä muutamia sanoja ymmärtävistä järjestelmistä moderneihin tekoälyä hyödyntäviin ratkaisuihin. Tämä teknologia on nykyään osana ihmisten jokapäiväistä arkea, sillä jokaisesta älypuhelimesta löytyy puheetunnistusteknologiaa, jota voidaan hyödyntää eri sovelluksien kanssa.

Suurimpia edistysaskeleita puheentunnistuksessa on ollut Markovin piilomalli (Hidden Markov Model, HMM). Sitä hyödyntävissä toteutuksissa käytetään tilastollista mallinnusta järjestelmässä, jonka oletetaan olevan Markov-prosessi. HMM:iä voidaan käyttää lukuisissa toteutuksissa puheentunnistuksen lisäksi, kuten signaalien prosessoinnissa tai osakemarkkinoiden ennustamisessa [2]. Nykyisellään syvät neuroverkot (Deep Neural Networks, DNNs) ovat suuren kiinnostuksen kohteena ja mahdollistavat entistä monimutkaisempia ja tehokkaampia ratkaisuja.

Puheentunnistuksen haasteena toimii erityisesti melu, jonka laatu ja voimakkuus vaihtelevat eri ympäristöjen mukaan. Puheentunnistus voi toimia täydellisesti hiljaisessa tilassa, mutta robottien toimintaympäristö on harvoin täysin meluton. Siksi on tärkeää kiinnittää huomiota melunsuodatukseen, johon voi hyödyntää erilaisia tekniikoita niin laitteisto- kuin ohjelmistopuolella.

Lisäksi puheentunnistus ei yksinään riitä hyvään kommunikaatioon vaan robotin tulee myös ymmärtää kuulemaansa. Tänä päivänä keskustelutekoälyt ovat kasvava trendi ja jopa yhteiskunnallisesti merkittävä aihe. OpenAI:n vuonna 2022 julkaiseman ChatGPT:n myötä tekoälyn käyttöä on jouduttu puimaan niin oppilaitoksissa kuin työpaikoillakin. Keskustelutekoälyjä on ollut jo kauan, mutta aiemmin ne ovat olleet pääasiassa yksinkertaisia sääntöpohjaisia palvelubotteja nettisivuilla. Keskustelutekoälyn valjastaminen puhetta tunnistavan robotin käyttöön antaa mahdollisuudet luontevaan keskusteluun ihmisen ja koneen välillä.

2. PUHEENTUNNISTUSJÄRJESTELMÄT

Tärkeä osa ihmisten ja robottien välistä vuorovaikutusta on toimiva puheentunnistus, joka mahdollistaa keskustelun robotin ja ihmisen välillä. Vuosikymmenien ajan kehitelty puheentunnistusteknologia on nykypäivänä jo hyvin vakaata ja laadukasta. Isoja harppauksia sen kehittymisessä ovat olleet Markovin piilomallin sekä syvien neuroverkkojen käyttöönotto. Puheentunnistusteknologia koostuu useasta eri osasta ja haasteita toimivan järjestelmän luomiseen tuo vaihtelevuus niin melun määrässä, puhutussa kielessä kuin itse puhujassa.

2.1. Modernit puheentunnistussovellukset

Puheentunnistusteknologiaa on kehitelty jo 1950-luvulta lähtien [3] ja nykyisin siitä on tullut osa ihmisten jokapäiväistä elämää. Yksi näkyvä esimerkki ovat modernit puheohjauksella toimivat virtuaaliavustajat, jotka voivat suorittaa erilaisia tehtäviä, kuten tiedonhakua tai musiikin toistamista. Lisäksi niiden kautta voi hallita myös kodin muita laitteita internetin välityksellä, jolloin puheohjauksella voi vaikkapa kontrolloida asunnon valaistusta. Useat suuryritykset, kuten Amazon, Apple ja Google ovat kehittäneet oman virtuaaliavustajansa, jotka kulkevat helposti taskussa mukana minne tahansa. Vuonna 2017 tehdyssä tutkimuksessa 46 prosenttia yhdysvaltalaisista aikuisista vastasi käyttävänsä virtuaaliavustajaa, suurin osa puhelimella [4].

Toinen esimerkki puheentunnistusteknologian käytöstä löytyy opetustoiminnasta. Puheentunnistusjärjestelmien avulla oppilaille voidaan opettaa oikeanlaista lausumista sekä lisätä puhumisen itsevarmuutta. Järjestelmät voivat arvioida ja jopa pyrkiä tunnistamaan virheitä lausumisessa [5]. Lisäksi puheentunnistuksella on lukuisia muita käyttökohteita, kuten jonotusaikojen lyhentäminen asiakaspalvelukeskuksissa, puhelimen käytöstä aiheutuvien riskien vähentäminen autoissa ja puhujan tunnistaminen turvajärjestelmissä. Nämä järjestelmät helpottavat ihmisten jokapäiväistä elämää, kehittävät turvallisuutta ja yleisesti lisäävät mukavuutta. Kuitenkin teknologian kehittyessä myös näihin järjestelmiin liittyy riskejä. Esimerkiksi turvajärjestelmissä puheentunnistusta voidaan väärinkäyttää nykyaikaisen puhesynteesin avulla, sillä tekoälyllä voidaan tuottaa täysin realistista puhetta lähes vaivatta. Tällöin puheentunnistusjärjestelmän on lähes mahdotonta havaita onko puhuja aito ihminen vai ei, mikä voi mahdollistaa väärinkäytöksiä esimerkiksi asiakkaan tunnistamisessa puhelimen välityksellä [6].

2.2. Sosiaaliset robotit

Sosiaaliset robotit ovat yleistyneet nopeasti viime vuosien aikana ja nykypäivänä ne ovatkin jo mukana esimerkiksi hoitotyössä [7, 8], opetuksessa [9, 10, 11] sekä asiakaspalvelutehtävissä [12]. Sosiaalisille roboteille nimensäkin mukaan on tärkeää kyky kommunikoida ihmisten kanssa. Puheentunnistus mahdollistaa sulavan kanssakäymisen ihmisen ja robotin välillä ja sen vuoksi sosiaaliset robotit soveltuvat monenlaisiin ympäristöihin ja käyttötarkoituksiin. Kun robotteihin myös yhdistetään puhesynteesi, on ihmisen ja robotin välinen kommunikointi jo lähes ihmismäistä.

Ihmiset käyttävät kommunikaatiossaan kuitenkin muutakin kuin puhetta. Eleet kasvoilla, käsillä ja muilla kehon osilla ovat suuri osa kanssakäymistä. Eleiden, kuten käsien liikkeiden käyttäminen tietokoneiden ohjaamiseen on ollut keskeinen tutkimusaihe kymmenien vuosien ajan. Niillä on esimerkiksi pyritty korvaamaan laitteita, kuten hiiri ja näppäimistö [13]. Tavoitteena on ymmärtää ja hyödyntää kaikkia ihmisen kehon eleitä.

Lisäksi eräs keskeinen tutkimusaihe ihmisen ja tietokoneen välisessä vuorovaikutuksessa on ollut tunteiden tunnistaminen. Tunnistus voidaan suorittaa sekä puheesta, että kasvojen piirteistä [14, 15]. Tämän teknologian avulla tietokoneet pystyvät yhä paremmin ymmärtämään ihmisten toimintaa.

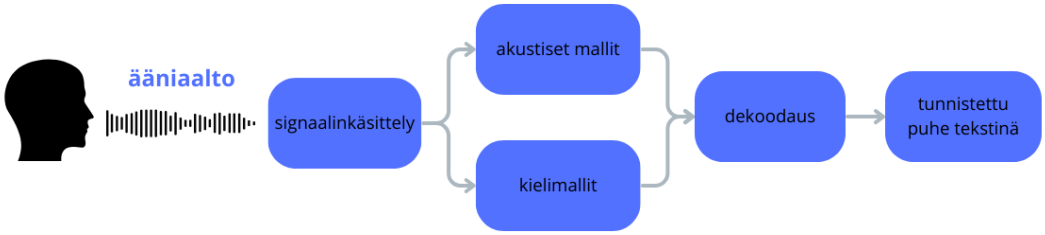
Lähes kymmenen vuotta sitten, vuonna 2014 SoftBank Robotics julkaisi Pepper-robotin, joka kykenee tunnistamaan erilaisia tunteita sekä äänensävyjä ja reagoimaan niihin [16]. Lisäksi se osaa havainnoida ympäristöään, liikkua, elehtiä ja käyttää kehonkieltä sekä käymään luontevia keskusteluja ihmisten kanssa. Pepperin jälkeen on tullut useita uusia vielä kehittyneempiä humanoidirobotteja, kuten Engineered Artsin Ameca-robotti [1] vuonna 2022. Lisäksi esimerkiksi Tesla on ilmoittanut kehittävänsä humanoidirobotti Optimusta [17].

2.3. Puheentunnistusjärjestelmän tekninen toteutus

Automaattisella puheentunnistuksella tarkoitetaan prosessia ja teknologiaa, jossa puhesignaali käännetään sitä vastaaviksi sanoiksi ja muiksi kielellisiksi kokonaisuuksiksi algoritmisesti laitteissa, tietokoneissa tai tietokoneklustereissa [18].

Puheentunnistuksen neljä päävaihetta ovat: 1. signaalinkäsittely, 2. akustiset mallit, 3. kielimallit sekä 4. dekoodaus [19, 20]. Kuvassa 1 on visualisoitu yksinkertaistettu perinteisen puheentunnistusjärjestelmän rakenne. Signaalinkäsittelyssä puhe muutetaan digitaaliseksi signaaliksi, jonka jälkeen sitä usein käsitellään laadun parantamiseksi esimerkiksi taustamelua vähentämällä. Jotta signaalista saadaan erotetuksi äänteitä, muutetaan se usein esimerkiksi Fourier-muunnoksen avulla taajuuksia kuvaaviksi spektrogrammeiksi. Akustisia malleja hyödyntämällä nämä äänteet voidaan yhdistää niitä vastaaviin kirjaimiin ja sanoihin. Ne eivät kuitenkaan yksinään riitä hyvään puheentunnistukseen vaan lisäksi tarvitaan kielimalleja, jotka arvioivat peräkkäisten sanojen todennäköisyyksiä lauseissa. Dekoodaus on viimeinen vaihe, jossa hakualgoritmi etsii kaikista todennäköisimmän sanajonon perustuen akustisten ja kielimallien tuloksiin.

[1] https://www.engineeredarts.co.uk/robot/ameca/



Kuva 1. Puheentunnistusjärjestelmän rakenne[2].

2.3.1. Signaalinkäsittely

Puheentunnistus perustuu äänteiden tunnistamiseen äänisignaalista. Tavallisesti signaalin taajuuskomponentit saadaan selville Fourier-muunnoksella, joka digitaalisessa signaalinkäsittelyssä tarkoittaa diskreettiä Fourier-muunnosta, tavallisimmin nopeaa Fourier-muunnosta (Fast Fourier Transform, FFT).

Fourier-analyysissä saadaan selville koko signaalin keston aikana esiintyvät taajuuskomponentit, mutta tarkat ajankohdat, milloin mikäkin taajuuskomponentti esiintyy, ei selviä tavallisella Fourier-analyysillä. Tavanomaista Fourier-muunnosta käytetään siksi pääasiassa staattisille signaaleille, jolloin taajuuskomponenttien oletusarvo pysyy samana koko signaalin keston ajan. Puheentunnistuksessa taajuuskomponenttien (äänteiden) ajallisella esiintymisellä on merkitystä, jolloin tavanomainen Fourier-analyysi ei ole riittävä.

Puhesignaalin tunnistukseen soveltuvia tekniikoita ovat Fourier-analyysin sovellus lyhytaikainen Fourier-muunnos (Short Time Fourier Transform, STFT). Lyhytaikaisen Fourier-muunnoksen idea on, että signaali jaotellaan ajalliselta kestoltaan hyvin lyhyisiin pätkiin ja näille tehdään erilliset Fourier-muunnokset. Näin voidaan rakentaa kolmiulotteinen esitys esimerkiksi spektrogrammin muodossa. [21]

Spektrogrammi on visuaalinen kuvaus siitä, miten signaalin taajuusspektri vaihtelee ajan funktiona. Spektrogrammeissa taajuus kuvataan tavallisesti pystyakselilla ja aika vaaka-akselilla. Signaalin amplitudi tai energia kuvataan koordinaattipisteen värin voimakkuutena.

2.3.2. Akustiset mallit

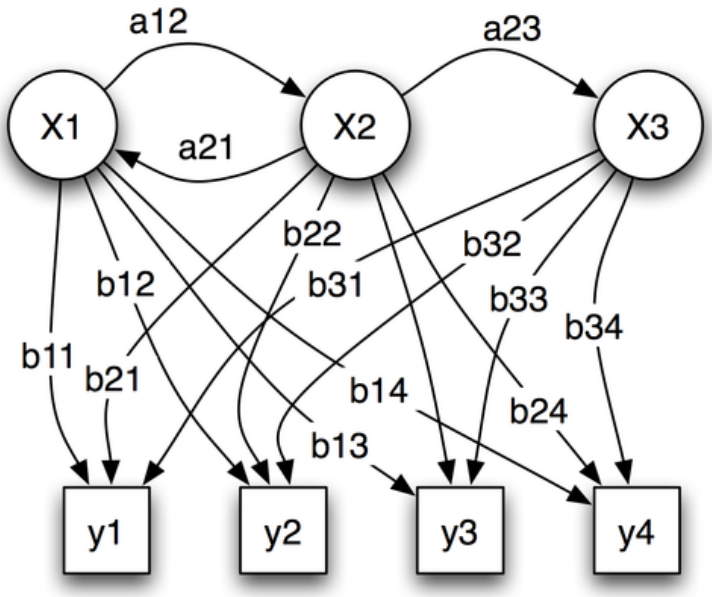
Akustiset mallit ovat keskeisiä osia puheentunnistusjärjestelmissä. Ne käsittelevät puhesignaalista saatua dataa ja yhdistelevät niitä kielen äänteisiin. Lisäksi ääntämysmallit ovat osa akustisia malleja [19]. Niiden avulla yksittäisistä peräkkäisistä äänteistä voidaan muodostaa järkeviä kokonaisia sanoja tai lauseita. Akustisessa mallinnuksessa luodaan tilastollinen esitys puhesignaalista eritellyille lyhyille taajuuksia kuvaaville piirrevektoreille, jotka voidaan eritellä esimerkiksi

[2]Kuva: Ida Haataja (lisenssi CC BY 4.0)

spektrogrammista. Piirrevektorit siis kuvaavat äänteiden taajuuksia ja niiden avulla voidaan määritellä äänteiden todennäköisyys kussakin signaalin osassa [22].

Äänteiden tunnistamisen haasteena on akustiikan vaihtelevuus erilaisten puhujien sekä ympäristön takia. Muun muassa puheen tyyli, nopeus, aksentti ja murteet sekä puhujan sukupuoli ja jopa stressi vaikuttavat tuloksen oikeellisuuteen [20]. Ympäristön melu ja muiden ihmisten taustapuhe tulisi myös pystyä erottamaan varsinaisesta puheesta, jota halutaan tunnistaa. Yleisimpiä akustisia malleja ovat muun muassa Markovin piilomalli (Hidden Markov Model, HMM) sekä syvät neuroverkot (Deep Neural Networks, DNNs). Nykyaikaiset puheentunnistusjärjestelmät ovat usein näiden kahden hybridimalleja.

Markovin piilomalli on perinteinen ja jo vuosikymmeniä käytössä ollut tilastollinen malli, jossa oletetaan, että sanoja vastaavat peräkkäiset piirrevektorit on muodostettu Markovin ketjulla [19]. Markovin ketju koostuu diskreeteistä tiloista ja seuraava tila riippuu vain ja ainoastaan nykyisestä tilasta. Markovin piilomallissa on kaksi tällaista ketjua, joista toinen on havaittavissa ja toinen on piilotettu [20]. Kuvassa 2 on visualisoitu, kuinka havaitut tilat saadaan piilotetuista tiloista eri todennäköisyyksillä. Puheentunnistusjärjestelmässä havaittavissa olevat tilat esittävät puheesta eriteltyjä piirrevektoreita ja piilotetut tilat taas äänteitä.



Kuva 2. Havainnekuva Markovin piilomallista. X:t kuvaavat piilotettuja tiloja, y:t havaittuja tiloja ja b:t ja a:t tilojen välisiä todennäköisyyksiä.[4].

Puheentunnistuksessa Markovin piilomalli perustuu usein Gaussin sekoitusmalliin (Gaussian Mixture Model – Hidden Markov Model, GMM-HMM), jolloin oletetaan, että jokainen havaittu tila on saatu Gaussin sekoitusmallin mukaisesti piilotetusta tilasta [20].

[4]Kuva: https://commons.wikimedia.org/wiki/File:HiddenMarkovModel.png (Lisenssi CC BY 3.0)

Syvät neuroverkot ovat monikerroksisia neuroverkkoja, joissa on useita piilotettuja kerroksia sisä- ja ulostulokerrosten lisäksi [20]. Useiden kerrosten ansiosta syvät neuroverkot voivat oppia tunnistamaan äänteitä hyvin monimutkaisista ja vaihtelevista puhesignaaleista. Syväoppiminen on parantanut puheentunnistuksen laatua huomattavasti ja nykyään useimmat järjestelmät hyödyntävät syviä neuroverkkoja niiden joustavuuden vuoksi. Kuitenkin luotettavan ja laadukkaan neuroverkon kehitys on monimutkainen ja aikaa vievä prosessi, joka vaatii todella suuria määriä koulutusdataa.

2.3.3. Kielimallit

Kielimallit toimivat akustisten mallien tukena ja etsivät kieleen perustuen todennäköisimmin sopivia sanoja lauseisiin. Ne usein perustuvat esimerkiksi tunnistettavan kielen kielioppiin ja siten arvioivat muodostettavien lauseiden todennäköisyyksiä ja järkevyyttä [22]. Kielestä muodostetaan tilastollinen malli, jota voidaan käyttää jo akustisten mallien yhteydessä. Siten jokaisen sanan esiintymiselle ja ääntämykselle voidaan laskea todennäköisyydet, jolloin saadaan järkeviä lauseita aikaan. Eräs tyypillisistä kielimalleista on n-gram-malli, joka arvioi sanan todennäköisyyttä perustuen n-1 edelliseen sanaan. Esimerkiksi 3-gramissa sanan todennäköisyys riippuu kahdesta edeltävästä sanasta. Malli vaatii miljoonia sanoja koulutusdataksi [19].

2.4. Melunsuodatus

Melunsuodatus on tärkeä osa toimivaa puheentunnistusjärjestelmää. Erilaiset häiriöt, taustamelu tai toisen ihmisen puhe tekee puheentunnistuksesta epätarkkaa tai toisinaan jopa mahdotonta. Lisäksi puhujan etäisyys mikrofoniin vaikuttaa puheentunnistuksen laatuun ja äänekkäässä ympäristössä sen vaikutus on suurempi kaikuelementtien vuoksi. Melunsuodatukseen voidaan vaikuttaa esimerkiksi sopivalla mikrofonilla sekä ohjelmallisesti erilaisten algoritmien avulla.

2.4.1. Mikrofonin vaikutus melunsuodatukseen

Erittäin tärkeä osa toimivan puheentunnistussovelluksen kehittämisessä on oikeanlaisen mikrofonin valinta. Sovellusympäristön häiriötekijöitä voidaan pyrkiä minimoimaan mikrofonin ominaisuuksilla. Mikrofonin laatu yleisesti vaikuttaa myös puheentunnistuksen laatuun, sillä huonolaatuisesta signaalista on vaikea erottaa eri sanoja ja äänteitä. On hyvä, jos mikrofonin vastaanottoherkkyyttä on mahdollista muuttaa, koska jo sillä taustamelun vaikutusta voidaan pienentää. Erot hiljaisen ja meluisan tilan välillä myös helpottuvat, jos asetuksia voi muuttaa olosuhteiden mukaan. Mikrofonien ryhmäprosessointi auttaa merkittävästi melunsuodatuksessa, kun puhesignaalista pyritään poistamaan mahdollisimman paljon häiriöitä, kuten taustakohinaa, muita melun lähteitä ja erilaisia kaikukomponentteja ennen varsinaista puheentunnistusta [23]. Mikrofoniryhmä (microphone array) koostuu useasta

strategisesti asetellusta mikrofonista, jotka vastaanottavat ääntä eri suunnista. Tällainen järjestelmä helpottaa äänen keilanmuodostusta eli hahmotusta siitä, mistä suunnasta tietty ääni tulee. Mitä enemmän mikrofoneja on, sitä paremmin keilanmuodostus toimii [24]. Tehokkaalla äänilähteen paikantamisella ja keilanmuodostuksella voidaan merkittävästi vähentää signaalin häiriökohinaa, sekä muita ylimääräisiä äänenlähteitä sekä heijastuvia, taittuvia ja siroavia kaikukomponentteja, jotka vaikeuttavat olennaisesti puheentunnistusta [25].

Kaikukomponenttien haasteet ja keilanmuodostuksen tärkeys korostuvat erityisesti, kun puhutaan kaukaisesta puheentunnistuksesta (Distant Speech Recognition, DSR), eli puheentunnistuksesta, jossa pyritään maksimoimaan puhujan ja vastaanottimen välinen etäisyys. Etäisyyden kasvaessa vastaanottavan mikrofonin ja äänenlähteen välillä, kasvaa myös eri suunnista heijastuvien, taittuvien ja siroavien äänikomponenttien määrä vastaanottavassa mikrofonissa. Tämä aiheuttaa useammasta mikrofonista yhdistettävän signaalin aaltomuodon vääristymistä ja vaihesiirtoa. Kaukaisen etäisyyden puheentunnistusjärjestelmät ovat kuitenkin selkeästi olleet viimeisimmän puheentunnistusta käsittelevän tutkimuksen erityisenä kiinnostuksen kohteena [26], [27]. Suurempi etäisyys vastaanottavista mikrofoneista vapauttaa käyttäjät monista käyttöliittymällisistä rajoitteista, kun mikrofonin ei tarvitse olla täysin lähettyvillä puheentunnistussovelluksia käytettäessä.

Kaukaisen puheentunnistuksen järjestelmä koostuu seuraavista komponenteista:

- puhujan sijainnin tunnistaminen,
- keilanmuodostus,
- jälkisuodatus ja
- varsinainen puheensuodatus

Puhujan sijainnin tunnistaminen tunnistaa sijainnin, josta puheen ääniaallot tulevat. Signaali välitetään keilanmuodostukselle, jossa mikrofonien vastaanotto voidaan kohdistaa parhaalla mahdollisella tavalla puhujan suuntaan signaali-kohina -suhdetta hyödyntäen. Vastaanoton kohdistaminen voidaan tarvittaessa suorittaa myös mekaanisesti kääntämällä mikrofonia. Luonteva ratkaisu tällaisessa tilanteessa olisi mikrofonin sijoittaminen esimerkiksi robotin päähän, jonka robotti kääntäisi puhujaa päin. Lopuksi kohdistetun signaalin laatua ja häiriöpuhtautta voidaan vielä parantaa jälkiprosessoinnilla ennen signaalin välittämistä varsinaiselle puheentunnistusyksikölle. Äänen suuntaamiseen on kolme erilaista tekniikkaa: 1) maksimaalisen vastaanottotehon tuottavan suunnan etsiminen, 2) lähteen paikallistaminen korkean resoluution spektraalisilla arviointitekniikoilla kuten esimerkiksi aliavaruuden algoritmeilla ja 3) arvioimalla sijainti signaalikomponenttien aikaviivästymistä mikrofoneissa. Yksinkertaisimpia ja laskennallisesti tehokkaimpia ovat aikaviivästyksen laskentaan perustuvat sovellukset [26]. Puheentunnistuksen onnistumiseen liittyy monia haasteita. Ensinnäkin puhujan sijainnin tunnistaminen voi epäonnistua, jolloin mikrofonien keilat eivät kohdistu puhujan suuntaan. Tällöin voi seurauksena olla puhesignaalin katkaiseminen taustameluna, koska suunnatun mikrofonisarjan keilan ulkopuoliset äänilähteet tulevat herkästi suodatetuiksi pois häiriööäninä. On myös mahdollista, että mikrofoniryhmän mikrofoneissa on laatupoikkeamia ja niillä on esimerkiksi toisistaan poikkeavia amplitudi- ja

vaihevasteita tai mikrofonit voi olla sijoiteltu piirilevylle virheellisiin paikkoihin. Kaikki nämä tekijät heikentävät mikrofonien keilan muodostusta. [26]

2.4.2. Melunsuodatus ohjelmallisesti

Melunsuodatusta voidaan tehdä myös ohjelmallisesti monin eri tavoin. Yksi keino on käyttää jo järjestelmän koulutusvaiheessa dataa, jossa on taustamelua, jolloin puheentunnistusjärjestelmä on jo ”tottunut” meluisaan puheeseen [28]. Puheen aktiivisuuden tunnistus (voice activity detection, VAD) osaa erotella signaalista ne hetket, jolloin puhetta esiintyy. Tällöin puheentunnistusjärjestelmälle voidaan lähettää vain puhetta sisältävät signaalit, jolloin puheentunnistusalgoritmin ei tarvitse yrittää tulkita taustamelua ja siten aiheuttaa vääriä tuloksia [29].

Meluisaa puhesignaalia voidaan myös itsessään yrittää selkeyttää erilaisten algoritmien avulla. Puheen parantamisen (speech enhancement) tarkoitus on poistaa ylimääräiset äänet sekä kohina puhesignaalista ja siten parantaa puheen laatua [30]. Keinoja tähän ovat muun muassa spektrivähennys (spectral substratction), jossa meluisa signaali muutetaan taajuusspektreiksi, jonka jälkeen niistä erotellaan puheen ja taustamelun komponentit [30]. Lopuksi meluisasta puhesignaalista vähennetään arvioidut melun taajuusspektrit, jolloin itse puhe on selkeämpää. Tämä toimii yleensä hyvin, jos taustamelu on tasaista ja ennustettavaa [31]. Toinen yleinen metodi on käyttää Wiener-suodatinta, joka minimoi keskimääräisen neliövirheen puheen ja melun välillä [32]. Viime vuosina syvät neuroverkot ovat myös tuottaneet erittäin hyviä tuloksia signaalin parantamisessa – tekoäly on mahdollista kouluttaa erottelemaan signaalista puhe ja taustamelu [33, 34]. Koska puheen erotus taustamelusta on hankala ja monimutkainen prosessi, neuroverkot sopivat erinomaisesti kyseiseen tehtävään.

2.5. Sovellusympäristö

Sovellusympäristöllä on suuri vaikutus puheentunnistuksen laatuun. Tässä työssä kehitettävän robotin pääasiallinen sovellusympäristö on Oulun yliopiston Hakijan päivä, joka järjestetään meluisassa ja kaikuisassa tilassa, jossa monet ihmiset tulevat puhumaan robotille. Yleisesti suurimman haitan puheentunnistukseen luovat taustamelu ja kaiku [35]. Sovellusympäristössä on näiden lisäksi muitakin haasteita. Käyttäjät tulevat puhumaan robotille yhtä aikaa, eri etäisyydeltä, eri voimakkuudella ja mahdollisesti jopa eri kielillä. Tämän lisäksi mahdollisista hakijoista etenkin tietoteknisesti lahjakkaat saattavat haluta testata robotin rajoja. Myöskään pahantahtoista testaamista ei voi jättää huomiotta ja se tulee huomioida koulutusdatan valinnassa. Sovellusympäristön aiheuttamia haittoja voi minimoida opastamalla käyttäjiä puhumaan yksi kerrallaan, tarpeeksi läheltä ja mahdollisimman selvällä suomen kielellä.

3. KESKUSTELUTEKOÄLYT

Keskustelutekoälyjen kehittyminen yksinkertaisesta 1960-luvulla luodusta yksittäisiä sanoja ja kuvioita tunnistavasta ELIZA-chatbotista moderneihin tekoälyä käyttäviin sovelluksiin on muuttanut ihmisten suhtautumista niin robotteihin kuin tekoälyyn yleensä. Parhaimmillaan tekoälyn kanssa käyty keskustelu voi vaikuttaa täysin luonnolliselta ihmisen kanssa käydyltä keskustelulta, mutta mitä itsevarmemmin chatbotit käyttäjilleen vastailevat, sen todennäköisemmin ne myös antavat täysin väärää, jopa vaarallista tietoa.

3.1. Ihmisen ja robotin välinen keskustelu

Keskustelurobotit ovat älykkäitä keskusteluagentteja, joiden avulla pyritään luomaan eri käyttöliittymiin ihmismäistä kanssakäymistä. Keskustelurobotteja voidaan käyttää moniin eri tarkoituksiin, kuten asiakaspalveluun, kyselyjen tekemiseen ja informaationhakuun. Ne voivat myös auttaa monimutkaisten ongelmien ratkaisemisessa ja tarjota asiantuntijatason neuvoja. Asiakaspalvelutilanteessa ne voivat vasta tarvittaessa ohjata asiakkaan ihmisen neuvottavaksi, helpottaen ihmisten taakkaa.

Keskustelurobotit ja niissä käytettävät teknologiat ovat saaneet lähiaikoina erittäin paljon huomiota. Tähän vaikuttaa monet syyt, joista tärkeimpiä ovat tekoälyn kehittyminen, luonnollisen kielen ymmärtämisen (NLU) kehittyminen ja universaalien keskustelualustojen yleistyminen [36].

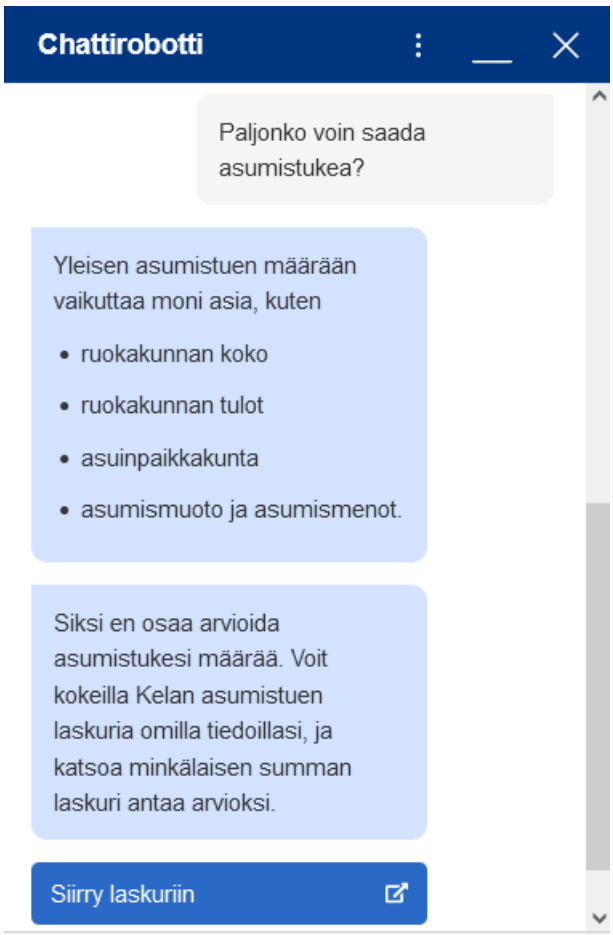
Keskusteluroboteilla on kaksi päätyyppiä: sääntöpohjaiset ja tekoälyyn perustuvat. Alun perin keskustelurobotit ovat olleet sääntöpohjaisia. Sääntöpohjaiset järjestelmät perustuvat tiukasti määriteltyihin sääntöihin, jotka kertovat robotille, miten sen tulee tiettyihin kysymyksiin vastata. Viime aikoina kuitenkin keskusteluroboteissa on voitu alkaa käyttämään tekoälyä. Tekoälyä käyttävät keskustelurobotit oppivat keskustelusta ja pystyvät tuottamaan luonnollisia viestejä.

3.2. Modernit keskustelutekoälyt

Tekoälyn kehittyessä puheentunnistusteknologia on myös ottanut suuria harppauksia eteenpäin ja robotit kykenevät keskustelemaan entistä enemmän ihmisen kaltaisesti. Nykypäivänä kuitenkin robottien sijaan on yleisempää löytää täysin virtuaalisia keskustelutekoälyjä. Esimerkiksi viime aikoina suurta huomiota on herättänyt OpenAI:n vuoden 2022 lopussa julkaisema ChatGPT. ChatGPT on neuroverkkopohjainen tekoäly, joka on osa suuria kielimalleja (Large Language Model, LLM). ChatGPT:n kanssa voi käydä melko luonnollisia keskusteluja niin englanniksi kuin suomeksikin – lukuisten muiden kielten ohella. Vaikka tekoälyn tuottama teksti on ajoittain hyvin vaikuttavaa, on se silti huomattavan epäluotettavaa ja kehitettävää on paljon.

Luotettavamman sisällön saamiseksi on parempi käyttää neuroverkkopohjaisen tekoälyn sijaan yksinkertaisempaa sääntöpohjaista teknologiaa, jonka vastauksia on helpompi hallita. Yleensä esimerkiksi nettisivujen palvelubotit on rakennettu

sääntöpohjaisen logiikan päälle, jotta virheellistä tietoa ei anneta käyttäjälle. Esimerkiksi Kelan nettisivuilla on suomenkielinen sääntöpohjainen "ChattiRobotti", joka vastaa yksinkertaisiin kysymyksiin antamalla linkit paikkoihin, joista vastaus voisi löytyä. Se ei siis suoraan anna vastausta, mutta esimerkiksi asumistuesta kysyttäessä kertoo, miksi ei voi vastata ja antaa linkin asumistuen laskuriin. Vaikeammissa kysymyksissä ChattiRobotti jättää vastaamatta kokonaan ja listaa parhaansa mukaan kolme linkkiä Kelan sivuille, joista vastaus voisi löytyä. Kelan ratkaisussa botti myös aluksi kertoo, minkälaisia kysymyksiä siltä kannattaa kysyä, ja varoittaa kirjoittamasta chattiin henkilötietoja. Kuvassa 3 on esimerkkikeskustelu Kelan Chattirobotin kanssa, jossa siltä kysytään asumistuen määrästä.



Kuva 3. Kelan Chattirobotti[5].

3.3. Vastauksen valinta

Keskustelurobotin vastauksen valinta on yksi toimivan keskustelurobotin suurimmista haasteista. Tarkoituksesta riippuen vastauksen valinta voidaan suorittaa eri tavalla. Keskustelurobotti voi antaa käyttäjälle valmiiksi kirjoitetun vastauksen tai luoda vastauksen automaattisesti.

[5]Kuva: `https://www.kela.fi/chattirobotti` (Lisenssi CC BY 4.0)

Yksinkertaisin mahdollinen tapa suorittaa vastauksen valinta keskustelurobotilla ovat sääntöpohjaiset mallit. Sääntöpohjaiset keskustelurobotit osaavat vastata ennalta määriteltyihin kysymyksiin. Myös vastaukset ovat erikseen määriteltyjä. Ne käyttävät usein AIML (Artificial Intelligence Markup Language) -kieltä, joka on XML-pohjainen keskustelurobottien ohjelmoimiseen käytettävä ohjelmointikieli. [37]

Tekoälyyn perustuvat generoivat mallit (Statistical Machine Translation Generative Models) koulutetaan opetusdatan avulla vastaamaan kysymyksiin luontevasti. Mallit oppivat datasta ja pystyvät muodostamaan vastauksen automaattisesti. Ne voivat verrata annettua syötettä lukuisiin muihin keskusteluihin, ja valitsevat sen perusteella parhaimman vastauksen [38].

3.4. Keskustelutekoälyjen vaarat

Keskustelutekoälyjen kehityttyä on ihmisten luottamus niiden antamaan tietoon kasvanut. Kuitenkin generatiivisen tekoälyn luomat vastaukset eivät ole suoraan mistään lähteestä ja tietyissä tilanteissa niiden vastaukset ovat yksinkertaisesti väärin, mikä on herättänyt myös kritiikkiä tekoälyä kohtaan. Ilmiötä, jossa tekoälyä hyödyntävä kielimalli antaa vääristettyä tai järjenvastaista tietoa, kutsutaan LLM-hallusinoinniksi [39]. Väärä tieto voi aiheuttaa suuria vaaroja, jos käyttäjät uskovat vastauksiin esimerkiksi terveyteen liittyvissä kysymyksissä. Nämä vaarat kiteytyvät British Journal of Psychiatry -lehdessä julkaistussa artikkelissa, jossa todetaan, että "A major problem with generative AI is that people who do not know the correct answer to a question will not be able to tell if an answer is wrong"[40].

Keskustelutekoälyjen käyttö jokapäiväisessä elämässä on helppoa ja tehokasta. Niitä voi käyttää tiedonhakuun, ostoslistan luomiseen tai vaikka lomamatkan suunnitteluun. Lisääntyneessä käytössä piilee kuitenkin myös suuri tietoturvariski. Keskustelutekoälyille annetun tiedon määrä on itsessään tietoturvariski, mutta kun niiltä aletaan kysymään terveydenhuoltoon, pankkitoimintaan tai vaikka salasanan luomiseen liittyviä kysymyksiä, luovat ne aivan uuden hyökkäysvektorin hakkereille ja muille pahantekijöille. Pahimmillaan tekoäly voi oppia sille annetuista henkilötiedoista [41].

Ihmislähtöisestä datasta oppivan tekoälyn laatu kärsii ihmisten ennakkoluuloisuudesta. Nämä ihmisiltä opitut ennakkoluulot liittyvät yleensä sukupuoleen, ihonväriin tai uskontoon [42]. Erityisesti sukupuolten välisessä ennakkoluuloisuudessa tekoälyt ovat herättäneet keskustelua. Nämä ennakkoluulot voivat aiheuttaa jopa erittäin epätasa-arvoista tekstiä sukupuolten välillä [43].

4. TOTEUTUS

Työssä toteutettiin InMoov-humanoidirobottiin[6] puheentunnistusjärjestelmä sekä keskustelutekoäly, joka voi käydä keskusteluja tietotekniikkaan ja sen opiskeluun liittyen. Robotin ohjelmoinnissa käytetään Robot Operating System 2:ta[7] (ROS2) ja Python-ohjelmointikieltä.

4.1. InMoov-robotti

Työn tarkoituksena on kehittää puheentunnistusjärjestelmä ja sen kanssa yhdessä toimiva keskustelulogiikka InMoov-humanoidirobottiin. InMoov on avoimen lähdekoodin 3d-tulostettava robotti. Robotin ohjelmointiin käytetään avoimen lähdekoodin ROS2:ta, joka hyödyntää solmuja (node), jotka kommunikoivat keskenään julkaisemalla ja vastaanottamalla viestejä liittyen tiettyyn aiheeseen (topic). Tähän työhön liittyvää robottia on kehitetty useiden aiempien kurssien aikana ja robotilla onkin jo ylävartalo, josta löytyy esimerkiksi kameralla varustetut silmät sekä liikkuvat käsivarret, sormet, suu ja pää. Robotti kykenee muun muassa seuraamaan ihmisiä katsellaan ja tekemään joitakin ihmismäisiä liikkeitä ja eleitä. Robottiin ei kuitenkaan tätä ennen ole kehitetty minkäänlaisia keskusteluominaisuuksia, kuten puheentunnistusta tai puhesynteesiä. Siispä kehitystyö on aloitettava täysin alusta puheentunnistuksen ja keskustelulogiikan kannalta. Robotin toivotaan kykenevän yksinkertaiseen keskusteluun ihmisten kanssa ja olisi suotavaa, jos keskustelun lomassa robotti elehtisi ihmismäisesti, kuten kääntäisi katseen puhujaan päin. Kuitenkin työn ensisijaisena tavoitteena on saada luotua toimiva itsenäinen puheentunnistusjärjestelmä sekä yksinkertainen keskustelulogiikka, jotka toimivat yhdessä.

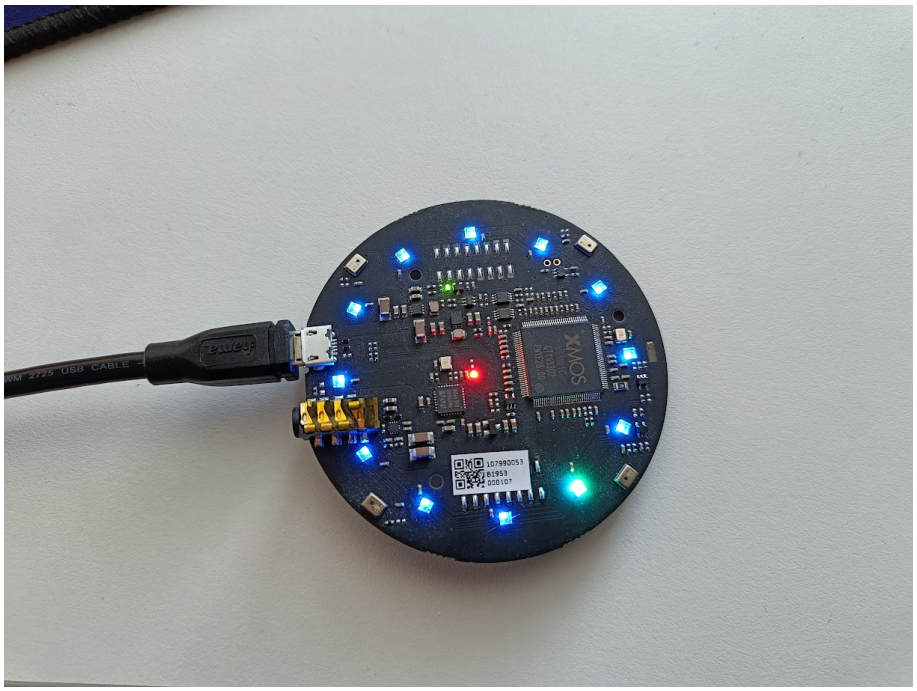
4.2. Puheentunnistusjärjestelmän ominaisuuksien valinta

Puheentunnistusjärjestelmälle asetettavat vaatimukset määräytyvät sovellusympäristön sekä kurssityöhön käytettävissä olevien resurssien ja ajan mukaisesti. Robotin on tarkoitus olla esillä Oulun yliopiston Hakijan päivien tietotekniikan standilla toimien vetonaulana tietotekniikan opiskeluista kiinnostuneille nuorille. Puheentunnistusjärjestelmän on kestettävä mahdollisimman hyvin tällaisia äänekkäitä messuolosuhteita, jotta se on kykenevä keskusteluun hakijoiden kanssa.

Mikrofoniksi valittiin kuvassa 4 näkyvä ReSpeaker Mic Array v.2.0.[8] Piirilevyyn sisältyy neljä ympyränmuotoisen levyn reunoille aseteltua mikrofonia. Äänen prosessoinnista vastaa XMOS XVF-3000[9] -prosessori, joka tukee monia melunsuodatuksen ja puheentunnistuksen teknologioita ja algoritmeja. Ominaisuuksiin kuuluu muun muassa VAD eli puheen aktiivisuuden tunnistus, äänen tulosuunnan tunnistus, keilanmuodostus, melunsuodatus sekä kaiunesto.

[6] https://inmoov.fr
[7] https://docs.ros.org/en/foxy/index.html
[8] https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0
[9] https://www.xmos.com/xvf3000

Mikrofoniryhmä pystyy havaitsemaan puhujan suunnan meluisassakin ympäristössä ja mikrofonin sisäänrakennettujen melunsuodatusominaisuuksien ansiosta se soveltuu hyvin työn tarkoituksiin.



Kuva 4. Työssä käytetty ReSpeaker Mic Array v.2.0. Vihreä led osoittaa suunnan, josta ääni tulee. [11]

Mikrofonin ajurit ovat saatavilla kaikille käytetyimmille tietokoneiden käyttöjärjestelmille ja mikrofonin konfigurointiin on käytettävissä helppokäyttöinen Python -skripti, jolla voidaan muun muassa valita mikrofonista ulos tulevien äänikanavien määrä. Mikrofoni on fyysisiltä mitoiltaan riittävän pienikokoinen sijoitettavaksi InMoov-humanoidirobottiin 3d-printattavilla soviteosilla.

Robotin puheentunnistusjärjestelmän on kyettävä tunnistamaan laaja valikoima sanoja suomen kielellä. Lisäksi olisi hyvä, jos robotti voisi ymmärtää myös useita muitakin kieliä. Kielimallien vaatimukset ovat korkeat, järjestelmän on tunnistettava esimerkiksi kysymykset ja toteamukset toisistaan, sen on ymmärrettävä erilaisia asiakokonaisuuksia ja käsitteitä. Näin ollen puheentunnistusjärjestelmäksi päädyttiin valitsemaan valmiiksi koulutettu puheentunnistusjärjestelmä. Saatavilla on mm. Google speech recognition API, Pocketsphinx ja Mozilla Deepspeech, joita jokaista kokeiltiin kehitystyön edetessä.

4.3. Puheentunnistusjärjestelmä

Puheentunnistusjärjestelmän toteutukseen työssä käytetään valmista Python-ohjelmointikielen SpeechRecognition -kirjastoa sen helppokäyttöisyyden ja joustavuuden takia. Kirjasto pystyy hyödyntämään useita eri API-rajapintoja

[11] Kuva: Ida Haataja (Lisenssi CC BY 4.0)

puheentunnistukseen. Työhön on valittu Googlen Web Speech -rajapinta, joka hyödyntää syviä neuroverkkoja ja pystyy tunnistamaan myös suomen kieltä. Käytännössä kirjaston ansiosta työssä on heti suomeksi toimiva puheentunnistustyökalu. Puheentunnistus tapahtuu omassa solmussaan, joka tunnistaa puheen ja lähettää sen sitten merkkijonona eteenpäin toisille solmuille - tässä tapauksessa keskustelulogiikasta vastaavalle Chatbot -solmulle. Puheentunnistussolmu aloittaa kuuntelun heti käynnistyessään ja pyrkii tunnistamaan puhetta Googlen puheentunnistusta käyttäen. Jos puhe tunnistetaan, tallennetaan se merkkijonoksi, joka lähetetään julkaistavaksi. Tunnistettu teksti myös tulostetaan näkyviin käyttäjän terminaaliin. Jos puhetta ei pystytä tunnistamaan, annetaan terminaaliin ilmoitus eikä viestiä lähetetä eteenpäin. Lisäksi solmu voi vastaanottaa viestejä muilta solmuilta aloittaakseen tai lopettaakseen kuuntelun.

4.4. Keskustelutekoäly

Tunnistetun puheen ymmärtämistä ja sille annettavaa vastausta varten luotiin oma solmu keskustelutekoälylle. Keskustelutekoälyn luomiseen hyödynnetään valmista sääntöpohjaista ChatterBot -kirjastoa eli ratkaisu ei hyödynnä tekoälyä. Kuitenkin valmiin ChatterBot -kirjaston ansiosta työssä on nopeasti valmis pohja, josta on helppo lähteä kehittämään toimivaa ohjelmaa.

ChatterBot sijaitsee omassa ChatBot -solmussa, jossa se vastaanottaa syötteitä puheentunnistussolmulta ja voi lähettää viestejä eteenpäin muun muassa puhesynteesin solmulle. ChatterBotille määritetään halutut logiikka-adapterit sekä oletusvastaus, jonka botti antaa ellei se löydä sopivaa vastausta sille koulutetusta datasta. Logiikka-adapterit ovat ChatterBot -keskustelutekoälyn tärkein osa. Ne määrittelevät, miten botti muodostaa vastauksen annettuun syötteeseen ja antavat sille luottamustason. Eri adapterien käyttö antaa botille kyvyn vastata erilaisiin kysymyksiin nopeasti ja luotettavasti - esimerkiksi matemaattisissa kysymyksissä botti voi hyödyntää erillistä matematiikka-adapteria, joka suorittaa laskutoimituksen sen sijaan, että se yrittäisi etsiä vastausta koulutusdatasta, jossa sillä tuskin olisi vastausta. Syötteen saadessaan botti siis alkaa käymään läpi eri logiikka-adaptereita, kunnes saa tarpeeksi hyvän luottamustason vastauksen. Jos tarpeeksi hyvää luottamustasoa ei saavuteta, siirrytään lopulta Low Confidence Response -adapteriin, joka antaa valmiiksi asetetun perusvastauksen. Botin luottamustasoa voidaan myös vaihdella - mitä alhaisempi taso on, sitä helpommin botti antaa vastauksen vähänkään sopivaan kysymykseen. Luottamustasolla on siis suuri vaikutus botin vastauksien laatuun. Työssä kokeiltiin eri luottamustasoja ja niiden vaikutusta vastauksen valintaan.

Ennen kuin botti osaa vastata kysymyksiin, se täytyy kouluttaa käyttäen erillistä 'train' -funktiota, jolle määritetään halutut koulutustiedostot. Koska osa datasta on englanninkielistä ja botin on tarkoitus puhua suomea, lisättiin ChatBot -solmuun myös Googlen Translate -ominaisuus, jolla suomenkieliset syötteet voidaan kääntää englanniksi. Siten botti voi vertailla käännettyä syötettä omiin englanninkielisiin kysymys-vastauspareihin. Jotta botti voi hyödyntää molempien kielien dataa, se etsii erilliset vastaukset suomenkieliseen sekä käännettyyn syötteeseen. Sen jälkeen se vertailee vastausten luottamustasoja, joista se lopulta valitsee luotettavamman.

4.4.1. Botin kouluttaminen

ChatterBot -kirjastossa on valmiita työkaluja keskustelutekoälyn kouluttamiseen. Se tarjoaa myös valmista englanninkielistä koulutusdataa. Kouluttamisen voi suorittaa antamalla botille joko kokonaisia keskusteluja tai kysymys-vastauspareja.

Jotta keskustelu esittelystandilla olisi luontevaa, robotin tulisi osata vastata kysymyksiin tietotekniikasta, sen opiskelusta sekä tietotekniikan alalla työskentelystä. Sovellusympäristöstä ja käyttäjistä johtuen sen olisi hyvä pystyä keskustelemaan myös muista asioista ja vastaamaan esimerkiksi tervehdyksiin.

Koska valmista suomenkielistä koulutusdataa ei ole paljoa ja robotin tulisi osata esitellä nimenomaan Oulun yliopiston tietotekniikan opintoja, sopivan koulutusdatan löytäminen on haastavaa. Siispä ratkaisuksi työhön päätettiin kirjoittaa tietotekniikan opiskeluun liittyvä koulutusdata itse ja muuten hyödyntää englanniksi löytyviä aineistoja, jotka käännetään suomeksi. Itsekirjoitettu data sisälsi yli sata kysymys-vastausparia liittyen itse tietotekniikkaan, sen opiskeluun sekä alalle hakemiseen ja työllistymiseen. Lisäksi omaa dataa kirjoitettiin myös vastaamaan kysymyksiin itse robottiin ja sen rakentamiseen liittyen. Näiden lisäksi kysymys-vastauspareja oli tervehdyksiä varten. Taulukkoon 1 on poimittu botille koulutettuja kysymys-vastauspareja eri aiheisiin liittyen.

Kysymys/syöte	Vastaus
Hei!	Hei, mitä kuuluu?
Millaisia taitoja tietotekniikan opiskelu kehittää?	Opiskellessa oppii monenlaisia taitoja riippuen siitä, mitä kursseja valitsee.
Mitä tietotekniikan opintoja Oulun yliopisto tarjoaa?	Oulun yliopisto tarjoaa tietotekniikan kandidaatin ja maisterin tutkinto-ohjelmia. Niihin kuuluu erilaisia kursseja ohjelmistotekniikkaan, tietoverkkojen suunnitteluun, tietoturvaan ja tekoälyyn liittyen.
Millaisia työpaikkoja on tietotekniikan alalla?	Tietotekniikan alalla on monia erilaisia työpaikkoja, kuten ohjelmistokehittäjä, tietokanta- ja järjestelmänhallintatehtävät, verkkoteknologiaan liittyvät tehtävät, tietoturva-asiantuntija ja IT-tuki.
Heippa!	Heippa, nähdään pian!

Taulukko 1. Botille koulutettuja kysymys-vastaus pareja.

Valmiin koulutusdatan puutteen lisäksi botin kouluttaminen isolla määrällä dataa osoittautui hitaaksi sillä ChatterBot -kirjasto ei sovellu hyvin isoon määrään koulutusdataa. Suurin valmis keskustelutekoälyn koulutusdata oli Microsoftin MSMARCO -aineisto, joka sisältää miljoona kysymys-vastausparia. Sen kouluttaminen ChatterBotilla olisi kestänyt liian kauan. Kuitenkaan bottia ei

tarvitse kouluttaa joka kerta uudelleen ja koulutetun botin voi lisätä versionhallintaan, joten isonkin datamäärän opettaminen olisi teoriassa mahdollista, jos aikaa on tarpeeksi.

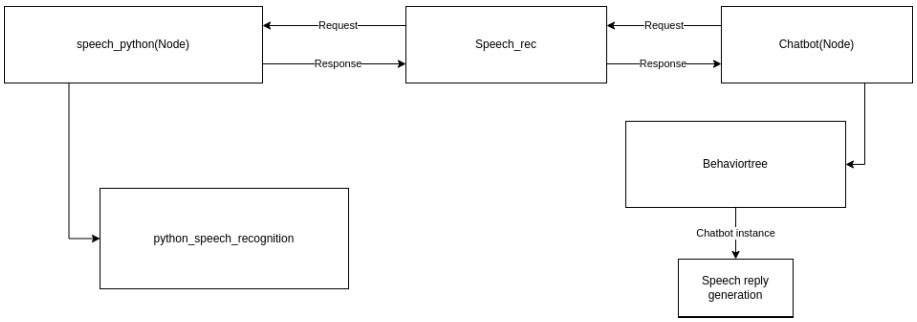
Botille koulutettiin mahdollisimman isoja määriä pääasiassa trivia-aiheisia kysymys-vastauspareja. Eri muodoissa löytyvän koulutusdatan kouluttamiseen loimme omat versiot ChatterBotin ChatterBotCorpusTrainer-funktiosta.

4.4.2. Esikäsittely

Esikäsittelemällä tunnistettua puhetta botin suorituskykyä voi parantaa. Esikäsittely tapahtuu ennen syötteen antamista logiikka-adapterille. Yksinkertaisimmat esikäsittelyfunktiot poistavat tekstistä kaiken ylimääräisen. Välimerkkien ja ylimääräisten välilyöntien poistaminen, pienaakkostus ja kääntäminen ovat yleisiä menetelmiä tekstin esikäsittelyyn. Tämän työn toteutuksessa syöte tulee puheentunnistuksesta, joten esimerkiksi oikeinkirjoituksen tarkistaminen ei ole tarpeellista. Työssämme tärkeä esikäsittelymetodi on kääntäminen, sillä käyttäessä englanninkielistä koulutusdataa, on syöte käännettävä ensin englanniksi. Esikäsittely ChatterBotilla on helppoa, joten koulutusvaiheessa testattiin eri esikäsittelymetodeja. Kuitenkin puheentunnistuksen ansiosta pelkkä kääntäminen oli tarpeellista.

4.5. Integraatio

Robottia varten puheentunnistusjärjestelmän ja keskustelutekoälyn täytyy olla yhteydessä toisiinsa. Työssä otettiin huomioon myös puhesynteesin käyttö, jotta robotti voi puhua vastauksena ääneen. Käytännössä puheentunnistusmoduuli kuuntelee käyttäjän puheen ja tunnistaa sen muodostaen merkkijonon. Merkkijono välitetään ChatBot-moduulille syötteenä, johon botti etsii vastauksen, jonka se julkaisee tekstinä. ChatBotin vastaus voidaan lähettää puhesynteesimoduulille, jolloin botin vastaus tulee varsinaisena äänenä ulos. Järjestelmän rakenne on havainnollistettu kuvassa 5. Puheentunnistus- ja ChatBotsolmu ajetaan omissa terminaaleissaan, jonka jälkeen ne ovat yhteydessä toisiinsa.



Kuva 5. Järjestelmän rakenne.

4.6. Testaus

Toteutuksen arvioimista varten suoritettiin yksinkertainen oikeaa keskustelutilannetta simuloiva testi hiljaisessa ja meluisassa ympäristössä. Yhteensä testejä tehtiin neljä kappaletta eri datamäärillä ja ChatterBotin luottamustasoilla. Järjestelmälle esitettiin yhdeksän eri syötettä, jonka jälkeen puheentunnistuksen onnistumista sekä botin vastauksien oikeellisuutta tarkasteltiin ja vertailtiin eri ympäristöjen suhteen. Testitilanteessa oli yksi puhuja, joka puhui mikrofoniin noin 40 senttimetrin päästä. Testit 1 ja 2 tehtiin hiljaisessa tilassa koko koulutusdatalla. Kuitenkin suurella koulutusdatan määrällä koulutettu botti osoittautui hyvin hitaaksi ja keskimäärin vastauksen antamisessa kesti 40-60 sekuntia. Niinpä testi 3 tehtiin samanlaisissa olosuhteissa, mutta botti koulutettiin vain omalla, tietotekniikan opintoihin liittyvällä datalla, jolloin botti antoi vastauksen lähes viiveettä. Testi 4 suoritettiin myös pienemmällä datamäärällä, mutta meluisassa tilassa yliopiston käytävällä. Botin luottamustaso oli 0.8 kaikissa testeissä paitsi toisessa testissä, jossa luottamustaso oli alempi 0.6.

Testauksessa botin kanssa käytiin keskustelu, johon kuului seitsemän kysymystä, tervehdys ja hyvästely. Vastauksia tarkasteltiin jakamalla ne hyväksyttäviin vastauksiin, huonoihin vastauksiin ja ei vastausta -perusvastaukseen. Taulukossa 2 on esitetty eri testitilanteet tiivistetysti. Eniten botilta saatiin hyväksyttäviä vastauksia (16 kappaletta) ja toiseksi eniten vastausta ei saatu (12 kappaletta). Näiden osuuksien perusteella botin koulutusdata on laadukasta, mutta sitä ei ole tarpeeksi.

	Melu, koulutusdata, luottamustaso	Tunnistus	Ok vastaus	Huono vastaus	Ei vastausta
Testi 1	Hiljainen, koko data, 0.8	9/9	4	2	3
Testi 2	Hiljainen, koko data, 0.6	9/9	5	2	2
Testi 3	Hiljainen, oma data, 0.8	9/9	3	2	4
Testi 4	Meluisa, oma data, 0.8	8/9	4	2	3

Taulukko 2. Testien asetelmat ja tulokset tiivistetysti.

Ensimmäisessä testissä hiljaisessa tilassa puheentunnistuksessa ei tullut virheitä, mutta botti jätti vastaamatta kolmeen syötteeseen. Pienentämällä luottamustasoa ja testaamalla aikaisemmilla syötteillä, saatiin yksi hyvä vastaus lisää. Testituloksista huomataan, kuinka pieni muutos voi vaikuttaa botin antamiin vastauksiin. Kaikista eniten botti jätti vastaamatta hiljaisessa tilassa omalla koulutusdatalla. Tästä voidaan päätellä, ettei koulutusdataa ollut riittävästi. Koska kysymykset olivat tietotekniikkaan liittyviä, ei koko koulutusdatan käyttäminen testaamisessa lisännyt huonojen vastausten määrää.

Taulukossa 3 on esitelty testeissä olleet kysymykset sekä puheentunnistusjärjestelmän suoriutuminen. Taulukosta nähdään, että puheentunnistusjärjestelmä toimi lähes täydellisesti meluisassakin tilassa. Kuitenkin melu aiheutti ongelmia silloin, kun järjestelmä alkoi tunnistamaan taustaääntä juuri ennen kuin testaaja alkoi puhumaan, jolloin puhe saattoi jäädä järjestelmältä kuulematta. Lisäksi järjestelmä tunnisti joitakin ylimääräisiä taustaääniä, joita se tulkitsi puheeksi. Ainoa väärin tunnistettu lause oli "mitä taiteen opiskeluun tarvitaan", kun botilta kysyttiin "mitä taitoja opiskeluun tarvitaan". Tässä tilanteessa kuitenkin havaittiin, ettei virheellinen puheentunnistus vaikuttanut keskustelutekoälyn antamaan vastaukseen, sillä botin vastaus oli sama kuin oikein tunnistetulla puheella.

Kysymykset	Testi 1	Testi 2	Testi 3	Testi 4
"Hei"	ok	ok	ok	ok
"Kuka olet?"	ok	ok	ok	ok
"Miten pääsen opiskelemaan tietotekniikkaa?"	ok	ok	ok	ok
"Millaista tietotekniikan opiskelu on?"	ok	ok	ok	ok
"Mitä taitoja opiskeluun tarvitaan?"	ok	ok	ok	x
"Millaisia töitä alalla on?"	ok	ok	ok	ok
"Kuinka paljon ohjelmointitaitoja tarvitaan?"	ok	ok	ok	ok
"Paljonko on 3 kertaa viisi?"	ok	ok	ok	ok
"Näkemiin"	ok	ok	ok	ok

Taulukko 3. Puheentunnistusjärjestelmän suoritus. "Ok"tarkoittaa, että kysymys kuultiin täsmälleen oikein ja "x"merkkaa virheellistä tunnistusta.

Kysymys nro	Testi 1	Testi 2	Testi 3	Testi 4
1	ok	ok	ok	ok
2	huono	huono	ok	ok
3	ei vastausta	ei vastausta	ok	huono
4	ok	ok	ok	ok
5	ok	ok	huono	ok
6	ei vastausta	ok	ei vastausta	ei vastausta
7	huono	huono	huono	ei vastausta
8	ok	ok	ei vastausta	ok
9	ei vastausta	ei vastausta	ei vastausta	ei vastausta

Taulukko 4. Botin vastauksien laatu kysymystä kohti.

Eri kysymysten välillä on myös paljon vaihtelua vastauksen laadun suhteen, mikä on selitettävissä puutteellisella koulutusdatalla. Taulukossa 4 on esitelty botin vastauksien laatu yksittäisiin kysymyksiin, josta eri kysymysten eroja on helppo havainnoida - joihinkin kysymyksiin botti vastasi joka kerta oikein, osassa vastaukset vaihtelivat ja esimerkiksi kysymykseen 9, "Näkemiin", botti ei antanut vastausta kertaakaan. Botin koulutusdatassa ei ollut kyseistä hyvästelyä tai vastaavaa sanaa, jonka vuoksi se ei osannut siihen vastata. Sen sijaan esimerkiksi alkutervehdykseen vastattiin jokaisella kerralla hyväksyttävästi.

Testatessa koko koulutusdatalla havaittiin myös tilanne, jossa botti vastaa kysymykseen väärin, mutta vastaus sattuu sopimaan kysymykseen. Kysyttäessä "kuka olet"botti vastasi englanninkielisestä koulutusdatasta löytyvään kysymykseen "What are young swans known as?". Vastaus "cygnets"olisi erittäin hyvä nimi robotille, vaikka vastaus onkin väärin. Tämä on listattu tuloksissa huonoksi vastaukseksi.

Testauksessa yhdeksi ongelmaksi osoittautui järjestelmän jumittuminen aika ajoin. Tähän voi vaikuttaa käytössä ollut suorituskyvyltään huono ja hidas tietokone eli mahdollisesti paremmalla laitteistolla ongelmaa ei esiintyisi. Kuitenkin jumittumisten takia järjestelmä piti käynnistää välillä uudestaan, mikä ei oikeassa käyttötilanteessa olisi suotavaa.

5. POHDINTA

Testitilanteissa havaittiin, että puheentunnistusjärjestelmä toimii melko luotettavasti. Hiljaisessa tilassa se tunnisti syötteet joka kerta oikein. Meluisassa tilassa se tunnisti yhden syötteen väärin yhdeksästä, minkä lisäksi se tunnisti joitain ylimääräisiä taustaääniä. On otettava huomioon, että oikeassa tilanteessa puheentunnistukseen vaikuttaa useat eri tekijät: mm. taustamelu, sen voimakkuus ja laatu, puhujan etäisyys mikrofonista, puheen äänenkorkeus, voimakkuus ja puhetyyli. Eri aspektien ja niiden yhdistelmien vaikutuksen määrittely on haastavaa ja siksi tähän työhön valittiin vain hyvin yksinkertainen testitilanne. Testauksessa ei myöskään kokeiltu, miten järjestelmä reagoisi, jos puhujia olisi useita tai jos he puhuisivat yhtä aikaa.

Keskustelutekoälyyn tekoon valittu ChatterBot -kirjasto osoittautui heti pullonkaulaksi. Suuren koulutusdatan kouluttaminen oli liian hidasta ja pienemmällä koulutusdatalla keskustelutekoälyn laatu jäi heikoksi. Vaihtoehtona oman keskustelutekoälyn kouluttamiselle alusta alkaen olisi valmiin toteutuksen käyttäminen esimerkiksi hyödyntäen tekoälyä ja ChatGPT:n kaltaista tuotetta. Tällä hetkellä botin osaaminen riippuu täysin sille opetetusta datasta. Dataa pitäisi olla runsaasti, jotta botti löytäisi vastauksia eri tavalla esitettyihin kysymyksiin. Datan kirjoittaminen itse on pidemmän päälle hyvin hidasta ja raskasta, koska saman kysymyksen voi esittää usealla eri tavalla ja mahdollisia kysymyksiä on lähes loputtomasti. Neuroverkoilla toimiva tekoäly voisi olla vaihtoehto, mutta sääntöpohjaisessa ratkaisussa etuna on vastauksien luotettavuus ja paikkansapitävyys - erityisesti, koska robotin tulisi tietää hyvin spesifistä aiheesta eli tietotekniikan opiskelusta juurikin Oulun yliopistossa. Lisäksi, kuten testitilanteessa havaittiin, sääntöpohjaisesta mallista voi olla hyötyä väärin tunnistetun syötteen kanssa, jolloin botti voi verrata väärin tulkittua kysymystä sen datasta löytyviin kysymyksiin. Sen sijaan vapaa neuroverkkoja hyödyntävä tekoäly voisi vastata väärään syötteeseen täysin arvaamattomalla tavalla. Toki, oikein ja hyvin rajatusti kouluttamalla tällaisenkin tekoälyn voisi saada toimimaan luotettavasti, jotta niin kutsuttua LLM-hallusinontia ei tapahdu.

Sääntöpohjaiselle botille koulutettavan datan täytyy olla laadukasta ja relevanttia. Työssä käytetty suuri määrä valmiita trivia-aiheisia kysymys-vastauspareja osoittautuivat lopulta turhiksi, sillä järjestelmälle pitäisi osata esittää juuri oikea kysymys oikeassa muodossa, jotta se osaisi vastata oikein. Esimerkkinä tästä on testitilanteiden ulkopuolella kysytty kysymys "Kuka oli Yhdysvaltojen ensimmäinen presidentti?". Botin koulutusdatassa ei ollut juuri kyseistä kysymystä, mutta se löysi lähimpänä olleen kysymyksen, joka oli "Oliko Abraham Lincoln Yhdysvaltojen ensimmäinen presidentti?"ja niinpä botti antoikin vastaukseksi alkuperäiseen kysymykseen "Ei". Lisäksi tällaisten tietovisakysymysten osaaminen voi olla hieman turhaa robotin käyttötarkoitusta ajatellen, varsinkin kun kysymysten aihealueet eivät ole välttämättä kovin tuttuja suomalaisille sillä alkuperäinen data oli englanniksi. Järjestelmän ja botin kyvykkyys riippuvat siis tällä hetkellä täysin koulutusdatan laadusta ja määrästä.

Loppujen lopuksi, tämän hetkinen toteutus on toimiva ja se kykenee yksinkertaiseen keskusteluun niin hiljaisessa kuin meluisammassakin ympäristössä. Paremmalla koulutusdatalla botti varmasti osaisi vastata järkevästi sille esitettyihin kysymyksiin

ja kykenisi toimimaan luotettavasti sille tarkoitetussa käytössä eli Oulun yliopiston Hakijan päivillä tietotekniikan standilla.

6. JATKOKEHITYS

Työssä saatiin luotua yksinkertaiset puheentunnistusjärjestelmä sekä keskustelutekoäly. Molemmat järjestelmät vaatisivat vielä kehitystä tulevaisuudessa. Puheentunnistusjärjestelmää varten melunsuodatusta tulisi parantaa, jotta robotin käyttö meluisassa ympäristössä onnistuu luotettavammin, erityisesti taustaäänien kitkemiseen tulisi keskittyä. Käyttämällä suuntaavaa mikrofonia, ja kuuntelemalla siten vain yhdestä suunasta tulevaa ääntä voitaisiin suodattaa pois merkittävä määrä muusta puheesta erityisesti erittäin meluisassa tilassa. Myös mikrofonin asettelua ja etäisyyttä puhujasta sekä robotin fyysistä sijaintia tulisi miettiä tarkkaan, jotta tunnistus onnistuisi parhaiten. Tärkeä toimenpide olisi myös rajata robotin kuunteluaktiivisuutta esimerkiksi tilakoneen avulla. Puheentunnistusjärjestelmä voisi olla toimettomassa tilassa, kunnes sille sanottaisiin tietty aloitusfraasi, kuten "Hei, robotti", jolloin se alkaisi kuuntelemaan puhetta aktiivisesti. Myös keskustelun aikana robotin tulisi lopettaa kuuntelu tarvittaessa, jotta se ehtii vastata kysymykseen kerrallaan, ja erityisesti ettei se puhesynteesin lisäyksen myötä tunnista omaa puhettaan.

Keskustelutekoäly tarvitsee vielä paljon kehitystä, jotta kommunikaatio robotin kanssa olisi sulavampaa. Nykyiseen toteutukseen voisi kehittää paremman keskustelulogiikan tai sen voisi korvata esimerkiksi täysin uudella neuroverkkoja hyödyntävällä ratkaisulla. Nykyisen kaltaisella sääntöpohjaisella toteutuksella koulutusdataa tulisi opettaa runsaasti lisää sekä monipuolistaa sitä. Erityisesti tulisi osata varautua Hakijan päivien kävijöiden kysymyksiin. Valmiin toteutuksen hyödyntäminen keskustelutekoälyyn olisi oman toteutuksen luomista tehokkaampi ratkaisu. Valmiita ratkaisuja on useita, mutta esimerkiksi käyttämällä maksullista OpenAI:n ChatGPT API -ohjelmointirajapintaa voitaisiin vähentää kouluttamiseen kuluvaa työtä ja aikaa, sekä lisätä keskustelutekoälyn suorituskykyä. Tällaisen tekoälyn käytöllä robotti voisi myös ymmärtää ja puhua useita eri kieliä.

Jotta robotin kanssa keskusteleminen olisi luontevaa, tulisi puheentunnistus, keskustelutekoäly, puhesynteesi sekä robotin liikkeet yhdistää toimivaksi kokonaisuudeksi. Olennaista olisi erityisesti puhesynteesin lisääminen, jotta keskustelutekoälyn luoma teksti saataisiin varsinaisena puheena ulos. Lisäksi robotin muita ominaisuuksia voisi hyödyntää tai yhdistellä puheentunnistuksen kanssa. Puheentunnistusta voitaisiin helpottaa esimerkiksi, jos robotti osaisi kameralla tunnistaa lähimmän henkilön ja kuunnella samasta suunnasta kuuluvaa puhetta. Ihmismäisyyden lisäämiseksi robotin olisi toivottavaa elehtiä puheen, kuuntelemisen ja taukojen aikana.

7. PROJEKTIN KUVAUS

Työ suoritettiin ryhmätyönä osana Sulautettujen ohjelmistojen projekti -kurssia. Työ aloitettiin vuoden 2023 tammikuussa taustatutkimuksella sekä taustakappaleiden kirjoittamisella ja saman kevään aikana tehtiin myös puheentunnistusjärjestelmän sekä keskustelutekoälyn toteutus. Syksyllä 2023 toteutukset viimeisteltiin ja testaus suoritettiin 2024 keväällä, jolloin myös kirjoitettiin loppuun tausta- ja toteutuskappaleet, analysoitiin tulokset ja viimeisteltiin työ loppuun. Projektiin käytetyt tunnit ovat eriteltynä taulukossa 5.

Työn kuvaus	Ajankohta	Ida Haataja	Veikko Romppainen	Juha-Matti Runtti
Taustatutkimus ja kirjoituksen aloitus	01/2023-03/2023	48	46	50
Käytännön toteutus	02/2023-05/2023	39	71	75
Kirjoitus, toteutuksen viimeistely	08/2023-03/2024	66	58	50
Testaus	03/2024-03/2024	10	10	10
Tulosten kirjoitus, tekstin viimeistely	03/2024-05/2024	34	20	5
Yhteensä		197	205	190

Taulukko 5. Projektin aikajana sekä käytetyt työtunnit.

8. YHTEENVETO

Työssä luotiin yksinkertainen meluisassakin tilassa toimiva puheentunnistus- ja keskustelutekoälyjärjestelmä. Työn toteutuksessa kyettiin hyödyntämään pitkälti valmiita ohjelmointikirjastoja, jotka mahdollistivat esimerkiksi laadukkaan puheentunnistuksen. Vaikka nykyaikaisella tekoälyllä varustetut keskustelubotit ovat vaikuttavia ja edistyksellisiä, on robottia tehtäessä otettava huomioon niiden epävarmuus ja vaarat. Yksinkertainen sääntöpohjainen botti kykenee riittävän suurella ja laadukkaalla datalla järkevään keskusteluun, ainakin jos aihe on hyvin rajattu kuten työn tapauksessa. Jatkokehitystä varten erityisesti koulutusdataan tulisi kiinnittää huomiota sekä keskustelujärjestelmän integrointiin robottiin.

Humanoidirobottien kanssa keskustelu ei oletettavasti tule mullistumaan suuresti lähivuosina ja tulevaisuus, jossa jokaisella on oma robotti kotona tekemässä kotitöitä, näyttää yhä melko kaukaiselta. Kuitenkin tämän hetken suuri kiinnostus tekoälyä kohtaan voi viedä kehitystä nopeastikin eteenpäin. Esimerkiksi ChatGPT:n myötä tekoäly on tullut valtaväestöllekin tutuksi tai jopa osaksi arkea tarjoten apua lähes mihin tahansa ongelmaan. Kasvaneen kiinnostuksen ja lisääntyneen tutkimuksen myötä tekoäly voisi muuttaa muotoaan puhelimen tai tietokoneen ruudulta osaksi fyysistä humanoidirobottia. Robottien kehittyminen luontevasti keskusteleviksi ja käyttäytyviksi, hyödyllisiksi apulaisiksi ja lopulta ihmismäisesti elehtiviksi humanoideiksi voikin tapahtua nopeammin kuin keskiverto kuluttaja osaa odottaa.

9. VIITTEET

[1] Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24, 2010.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[3] Douglas O'Shaughnessy. Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41(10):2965–2979, 2008.

[4] Nearly half of americans use digital voice assistants, mostly on their smartphones. Pew Research Center URL: `https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/ft_17-12-07_voiceasst_users/`, 2017.

[5] Ambra Neri, Catia Cucchiarini, and Helmer Strik. Automatic speech recognition for second language learning: How and why it actually works. In *Proc. ICPhS*, pages 1157–1160, 2003.

[6] Joseph Cox. How i broke into a bank account with an ai-generated voice. *Vice. February*, 23:2023, 2023.

[7] Cory D Kidd, Will Taggart, and Sherry Turkle. A sociable robot to encourage social interaction among the elderly. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 3972–3976. IEEE, 2006.

[8] John-John Cabibihan, Hifza Javed, Marcelo Ang, and Sharifah Mariam Aljunied. Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. *International journal of social robotics*, 5:593–618, 2013.

[9] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[10] Sara Ekström and Lena Pareto. The dual role of humanoid robots in education: As didactic tools and social actors. *Education and information technologies*, 27(9):12609–12644, 2022.

[11] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science robotics*, 3(21):eaat5954, 2018.

[12] Youngjoon Choi, Miju Choi, Munhyang Oh, and Seongseop Kim. Service robots in hotels: understanding the service quality perceptions of human-robot

interaction. *Journal of Hospitality Marketing & Management*, 29(6):613–635, 2020.

[13] Pallavi Halarnkar, Sahil Shah, Harsh Shah, Hardik Shah, and Jay Shah. Gesture recognition technology: a review. *International Journal of Engineering Science and Technology*, 4(11):4648–4654, 2012.

[14] Fernando Alonso-Martin, Maria Malfaz, Joao Sequeira, Javier F Gorostiza, and Miguel A Salichs. A multimodal emotion detection system during human–robot interaction. *Sensors*, 13(11):15549–15581, 2013.

[15] Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, Naveen Chilamkurti, and Vishaka Nanayakkara. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimedia Tools and Applications*, 81(24):35173–35194, 2022.

[16] Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48, 2018.

[17] Yihao Su. Artificial intelligence: The significance of tesla bot. *Highlights in Science, Engineering and Technology*, 39:1351–1355, 2023.

[18] Li Deng and Douglas O'Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[19] Xuedong Huang and Li Deng. An overview of modern speech recognition. *Handbook of natural language processing*, 2:339–366, 2010.

[20] Dong Yu and Li Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

[21] Singh J. K. Gupta V. K. Rathore S. Tiwari M. Khare A. Aggarwal, R. Noise reduction of speech signal using wavelet transform with modified universal threshold. *International Journal of Computer Applications*, 20(5):14–19, 2011. DOI: http://dx.doi.org/10.5120/2431-3269.

[22] Mikko Kurimo. Puheentunnistus. *Puhe ja kieli*, (2):73–83, 2008.

[23] Michael L Seltzer. *Microphone array processing for robust speech recognition*. PhD thesis, Carnegie Mellon University, 2003.

[24] Mohammad Ebrahim Sadeghi, Hamid Sheikhzadeh, and Mohammad Javad Emadi. Speech improvement in noisy reverberant environments using virtual microphones along with proposed array geometry. *EURASIP Journal on Advances in Signal Processing*, 2022(1):1–20, 2022.

[25] Wangyou Zhang, Xuankai Chang, Christoph Boeddeker, Tomohiro Nakatani, Shinji Watanabe, and Yanmin Qian. End-to-end dereverberation, beamforming, and speech recognition in a cocktail party. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3173–3188, 2022.

[26] Kenichi Kumatani, John McDonough, and Bhiksha Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine*, 29(6):127–140, 2012.

[27] James Fosburgh, Dushyant Sharma, and Patrick A Naylor. Room adaptation of training data for distant speech recognition. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 71–75. IEEE, 2023.

[28] Ashok Kumar and Vikas Mittal. Hindi speech recognition in noisy environment using hybrid technique. *International Journal of Information Technology*, 13:483–492, 2021.

[29] Javier Ramirez, Juan Manuel Górriz, and José Carlos Segura. Voice activity detection. fundamentals and speech recognition system robustness. *Robust speech recognition and understanding*, 6(9):1–22, 2007.

[30] Nabanita Das, Sayan Chakraborty, Jyotismita Chaki, Neelamadhab Padhy, and Nilanjan Dey. Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology*, 24:883–901, 2021.

[31] Ekaterina Verteletskaya and Boris Simak. Noise reduction based on modified spectral subtraction method. *IAENG International journal of computer science*, 38(1):82–88, 2011.

[32] Rafael Attili Chiea, Márcio Holsbach Costa, and Guillaume Barrault. New insights on the optimality of parameterized wiener filters for speech enhancement applications. *Speech Communication*, 109:46–54, 2019.

[33] Nasir Saleem and Muhammad Irfan Khattak. Deep neural networks for speech enhancement in complex-noisy environments. 2020.

[34] Wissam A Jassim and Naomi Harte. Comparison of discrete transforms for deep-neural-networks-based speech enhancement. *IET Signal Processing*, 16(4):438–448, 2022.

[35] Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28, 2018.

[36] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[37] Jagdish Singh, Minnu Helen Joesph, and Khurshid Begum Abdul Jabbar. Rule-based chabot for student enquiries. In *Journal of Physics: Conference Series*, volume 1228, page 012060. IOP Publishing, 2019.

[38] Jack Cahn. Chatbot: Architecture, design, & development. *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*, 2017.

[39] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[40] Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35, 2024.

[41] Martin Hasal, Jana Nowaková, Khalifa Ahmed Saghair, Hussam Abdulla, Václav Snášel, and Lidia Ogiela. Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation: Practice and Experience*, 33(19):e6426, 2021.

[42] Daniel James Fuchs. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1):1, 2018.

[43] Nicole Gross. What chatgpt tells us about gender: a cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8):435, 2023.