EN TALASSISTENTS STRUKTUR OCH FUNKTION

Hur optimera en talassistent för åldringar?

Hanna Liman 1901240
Kandidatavhandling i datateknik
Handledare: Marina Walden
Fakultetet för naturvetenskaper och teknik
Åbo Akademi
Våren 2022

Hanna Liman
hanna.liman@abo.fi

Innehållsförteckning

Förteckning över förkortningar

ASR	*Taligenkänning* (Automatic speech recognition) – igenkänning av talade ord och meningar med användning av datorprogram
NLU	*Förståelse av naturligt språk* (Natural language understanding) – ett program som kan känna igen ord som skrivs eller sägs
DM	*Dialoghanterare* (Dialogue manager) – upprätthåller dialogflödet, hämtar information från databasen samt producerar svar till användaren.
KDB	*Kunskapsdatabas* (Knowledge database)
NLG	*Textgenerering* (Natural language generator) – generering av text från en beskrivning av vad som skall uttryckas
TTS	*Talsyntes* eller *text-till-tal* (Text to speech synthesis) – framställande av konstgjort mänskligt tal med datorteknik
GMM	*Gaussisk blandningsmodell* (Gaussian mixture model) – en modell som räknar ut sannolikhet genom normalfördelning
HMM	*Dold Markovmodell* (Hidden Markov model) – en modell som observerar tillstånd, dolda tllstånd samt övergångar mellan tillstånd
DNN	*Djupa neurala nätverk* (Deep neural networks) – är ett konstgjort neuralt nätverk med flera dolda lager
RNN	*Återkommande neurala nätverk* (Recurrent Neural Networks) – är ett neuralt nätverk därförbindelser mellan noder skapar en riktad graf
IC	Svensk översätnning (Intent Classification)
SF	Svensk översättning (Slot Filling)
SPSS	Programvara för statistisk analys (Statistical parameter speech synthesis)

Work in progress

1. Inledning

Användningen av internet i människors vardag har ökat under de senaste åren. Tekniken har utvecklats mycket och även tillgängligheten till internet därtill. Barn lär sig redan från ung ålder att använda mobiltelefoner och vuxna har möjligheten att arbeta hemifrån tack vare bland annat bärbara datorer. Äldre människor har dock inte alltid det lika lätt med tekniken. Många äldre i dag har vuxit upp utan internet och mobiltelefoner som gör det svårt för dem att använda dem. Dessutom lider många äldre av fysiska svårigheter som försvårar användningen av teknik.

Under år 2020 använde 42 procent av 75-89-åringar internet dagligen, medan endast 27 procent av 75-89-åringar gjorde det år 2013 [[1], [2]]. Fastän äldres användningen av internet nästan fördubblats sedan 2013 är antalet ännu väldigt lågt då man jämför med att hela 93 procent av Finlands befolkning använder internet dagligen. Men vad beror det på?

Många äldre lider av bland annat dåligt minne samt svårigheter att använda händerna som gör det svårt för dem att till exempel använda en smarttelefon. Synen kanske också är dålig, varpå det kan vara svårt att se något på en liten skärm. Eftersom internet är väldigt tillgängligt är det frågan om problem inom användargränssnittet. Vad är det optimala användargränssnittet så att åldringar lättare kunde använda bland annat internet?

Denna avhandling behandlar hur talassistenter fungerar samt hur en talassistent utvecklad specifikt för åldringar kunde se ut. Avhandlingen behandlar även grundligare olika komponenter som används i talassistenter. Därtill analyseras även utmaningar samt nyttiga aspekter som talassistenter för åldringar skulle bemöta.

Syftet med denna avhandling är att undersöka hur talassistenter fungerar och är uppbyggda samt undersöka vad som borde göras annorlunda då man utvecklar

talassistenter specifikt för åldringar. Tanken är att få insikt i varför talassistenter utvecklade för åldringar skulle vara bra samt vilka utmaningar det kunde medföra. Avhandlingen behandlar även hur användargränssnittet kunde se ut.

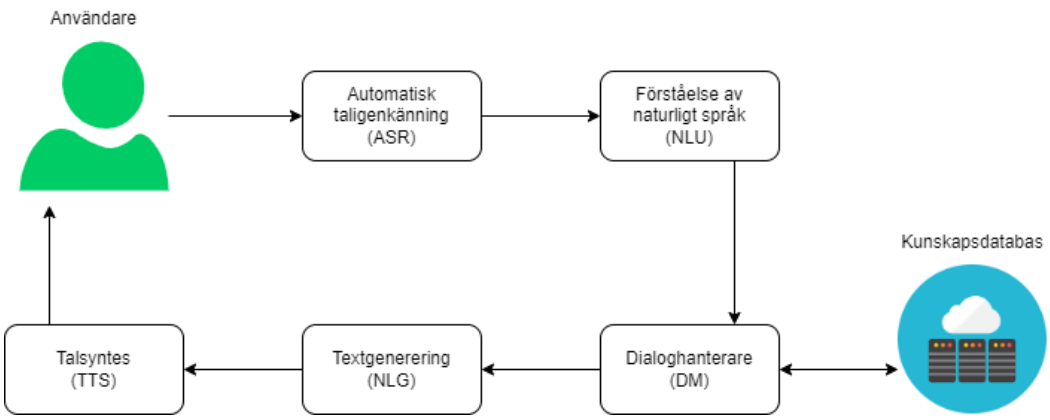
- Jag kommer också nämna om hur svårt det är för dagens äldre att lära sig ny teknik
- Jag kommer ännu skrva här om hur arthritic, dålig syn och dålig hörsel skulle påverka

2. Vad är en talassistent?

Work in progress

3. Hur ser strukturen ut?

Strukturen för talassistenter varierar beroende på vilka komponenter som används i systemet. Strukturen för talassistenter som denna avhandling kommer behandla inkluderar sex olika komponenter: *automatisk taligenkänning* (ASR), *förståelse av naturligt språk* (NLU), en *talhanterare* (DM), en *kunskapsdatabas* (KDB), *textgenerering* (NLU) och *talsyntes* (TTS) [[3], [4]]. Strukturen visualiserad även i figur 1 nedan.



Figur 1. Strukturen för en talassistent. Omarbetad och översatt från [[3]*, s. 4]*

Somliga talassistenter innehåller komponenten *förståelse av talat språk* (Spoken Language Understanding, SLU) istället för *förståelse av naturligt språk* (NLU), men båda komponenterna uppnår samma mål. Bland annat har företaget Amazon implementerat sin talassistent, *Amazon Alexa*, med komponenten NLU istället för SLU.

Komponenterna visualiserade i figur 1 behandlas nogrannare i följande sektion, där man går djupare in på vilka olika modeller och metoder som används i varje komponent samt hur varje komponent fungerar.

3.1 Automatisk taligenkänning (ASR)

Automatisk taligenkänning är en process som omvandlar tal till text med hjälp av algoritmer i en dator eller i ett datorkluster [5]. Genom att omvandla analoga signaler till digitala signaler kan en dator med hjälp av automatisk taligenkänning bilda ord och meningar. Komponenten Automatisk taligenkänning i talassistenter är oftast uppbygd av följande modeller: *särdragsextraktion* (feature extraction), en *akustisk modell* (acoustic model), en *uttalsmodell* (pronunciation model), en *språkmodell* (language model) och *textnormalisering* [[6], [7]].

Särdragsextraktion är den viktigaste delen inom automatisk taligenkänning. Särdragsextraktion samplar den analoga signalerna genom parametrisering för att omvandla signalen till en digital signal bestående av olika fonem. Parametriseringsmetoder som kan användas är bland annat: Linear Prediction Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC) och Linear Prediction Cepstral Coefficient (LPCC) [8]. Av dessa är metoden MFCC den vanligaste parametriseringsmetoden inom taligenkänning för att skilja på olika fonem [9].

Den akustiska modellen används efter särdragsextraktion för att producera en fonemsekvens genom att klarlägga hur sannolikt ett fonem följer efter ett annat fonem. Sannolikheten räknas med hjälp av olika modeller som den *Gaussiska*

blandningsmodellen (Gaussian mixture model, GMM) och den *Dolda Markovmodellen* (Hidden Markov model, HMM) [5].

Sedan år 2010 och uppkomsten av stordata (Big Data) har djupa neurala nätverk (Deep neural networks, DNN) även introducerats i samband med automatisk taligenkänning [5]. Djupa neurala nätverk fungerar med hjälp av djupinlärning där den akustiska modellen lär sig känna igen tal med hjälp av stora mängder träningsdata. Som följd av stordataperioden har användningen av stora mängder träningsdata möjliggjorts, varpå implementeringen av djupa neurala nätverk i akustiska modeller blivit lönsamt efter år 2010. Även ökad mängd forskning inom djupinlärning under det senaste decenniet har ökat användningen av neurala nätverk i ASR system.

Förutom att använda djupa neurala nätverk och modeller för att få fram fonemsekvensen, så kan den även räknas ut genom en kombination av neurala nätverk och HMM. Kombinationen av djupa neurala nätverk och HMM i den akustiska modellen har även visat sig vara bättre än en kombination av GMM och HMM [5]. Djupa neurala nätverk har visat sig ha bättre presision, vid tolkning av tal, än system som ändast använder sig av sannolikhetsberäkningar. God precision kan skapas genom att träna det djupa neurala nätverket med både varierande träningsdata samt träningsdata innehållande olika bakgrundsljud och brus.

Efter att den sannolika fonemfrekvensen blivit uträknad så jämförs sekvensen, med hjälp av uttalsmodellen, med ord som finns i språkmodellen [6]. Eftersom människor kan uttala ord på olika sätt, beroende på till exempel dialekter, så är det viktigt att ha en uttalsmodell för att minska på tolkningsfel av talet. Uttalsmodellen kollar vilket ord från språkmodellen som stämmer bäst överens med fonemsekvensen.

- Skall ännu skriva om textnormalisering i detta stycke
- En bild över en RNN språkmodell kommer läggas till ännu här? Nån bild i alla fall

3.2 Förståelse av naturligt språk (NLU)

Förståelse av naturligt språk (Natural Language Understanding, NLU) är en process där texten som skapats med hjälp av automatisk taligenkänning analyseras för att förstå vad texten betyder. NLU analyserar texten huvudsakligen för att ta reda på textens avsikt (Intent Classification, IC) samt för att ta reda på textens specifika entiteter (Slot Filling, SF) som krävs för att hämta läplig information [4]. Om till exempel texten som NLU analyserar är ”Hur långt är det från Helsingfors till Åbo?”, så identifieras ”långt” som avsikten med hjälp av IC och orden ”Helsingfors” och ”Åbo” identifieras däremot som specifika entiteter med hjälp av SF.

De strörsta utmaningarna som NLU kan stöta på är fel i texten som skall analyseras. Indatatexten som producerats med hjälp av automatisk taligenkänning kan i vissa fall vara fel på grund av att vissa ord eller sekvenser av ord låter samma. Bland annat homofoner, det vill säga ord som uttalas likadant men som har annan stavning och betydelse [10], kan bli fel vid automatiska taligenkänningen. Dessa fel kan uppstå fastän akustisk modellering gjorts för att bilda den mäst sannolika ordsekvensen. Till exempel frågan ”Kan jag köpa kål från K-market?” kan tolkas av talassistenten som ”Kan jag köpa kol från K-market?”, varpå innebörden av texten blir alldeles fel.

- Lösningar till utmaningarna gällande tolkningsfelen ovan kommer ännu behandlas här (note to self: källa 21)

3.3 Dialoghanterare (DM)

Dialoghanteraren är den centrala komponenten i en talassistent och mottar indata som skapats genom de tidigare processerna: automatisk taligenkänning och förståelse av naturligt språk. Dialoghanterarens uppgift är att hämta

information från en databas på basen av indata, producera meddelandet som skall framföras till användaren, samt upprätthålla dialogflödet [[11], [12]]. Dessa uppgifter kan dialoghanteraren genomföra med hjälp av dialogmodellering (Dialogue modeling) och dialogkontroll (Dialogue control).

Genom dialogmodellering sorteras indata för att registrera kommandon och frågor som sagts under tidigare dialog. Den sorterade dialogen sparas som dialoghistorik för att hålla reda på tillståndet i dialogen. Dialogkontrollen utförs sedan efter indata sorterats. Dialogkontrollen bestämmer vad dialoghanteraren skall göra till nästa på basen av tillståndet: antingen hämta information från databasen och framföra meddelandet, fråga efter mera indata eller kontrollera indatahistoriken för att kontrollera tidigare dialoger. Om till exempel användaren först frågat ”Vad heter finlands president?”, och fått svar på frågan, samt sedan ställt frågan ”Hur gammal är han?”, så vet dialoghanteraren att användaren frågar om presidentens ålder tack vare dialoghistoriken.

Dialogkontrollen som bestämmer hur dialogen skall fortsätta kan antingen vara systemstyrd (system-directed dialogue), användarstyrd (user-directed dialogue) eller ha mixad styrning (mixed-initiative dialogue) [12]. Dialogen nedan visar hur en systemstyrd dialog kan se ut:

Talassistent: *Vad heter du?*

Användare: *Hanna Liman*

Talassistent: *Hur gammal är du?*

Användare: *22 år.*

Talassistent: *Varifrån är du?*

Användare: *Jag är från Sibbo.*

Dialoghanterare i talassistenter har oftast en användarstyrd eller mixad dialogkontroll. Dialogen nedan som exempel på en användarstyrd dialogkontroll:

Användare: *Vad skall det vara för väder imorgon?*

Talassistent: *Det skall regna samt vara mulet.*

Användare: *Hur många grader har de lovat?*

Talassistent: *Det skall vara mellan 4 och 7 grader varmt.*

Dialoghanteraren strävar även efter att upprätthålla dialogflödet fastän tolkningsfel råkat ske i någon tidigare komponent [11]. Missförstånd uppstår även i dialoger mellan människor, varpå dialoghanteraren skapar ett mer naturlig dialog genom att kunna hantera tolkningsfel.

- Ännu exempel på va ett tolkningefel kan va från (källa 20)

3.4 Kunskapsdatabas (KDB)

Kommunikation med en kunskapsdatabas krävs för att en talassistent skall kunna hämta information samt svara på användarens frågor. Osv...

Varifrån hämtar till exempel Amazon Alexa information? Både Amazon Alexa och Microsofts Cortana använder sökmotorn Bing för att svara på frågor som användaren ställer [13].

Work in progress

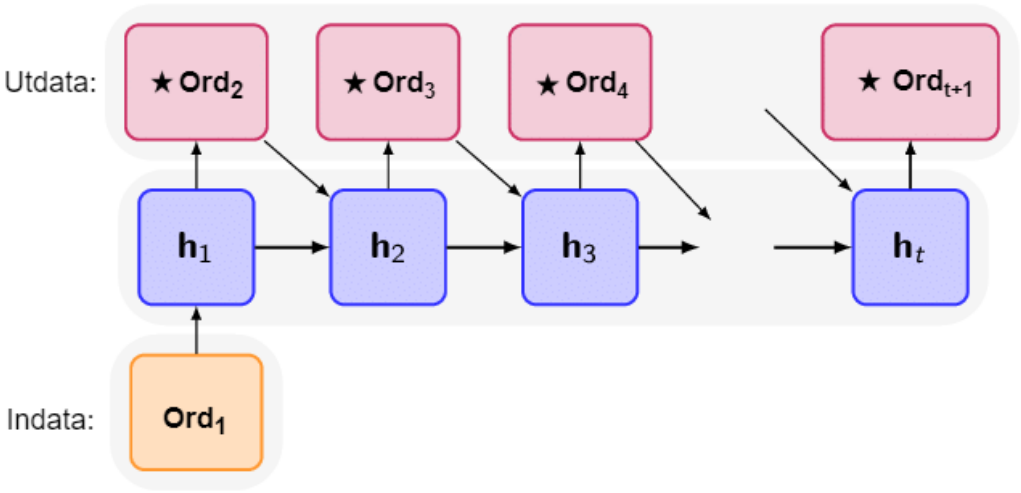
3.5 Textgenerering (NLG)

Textgenerering innebär generering av text från data som dialoghanteraren producerat [14]. Målet med textgenerering är att producera en så naturlig och flytande text som möjligt genom sammanhängande meningar. Användningen av textgenerering i talassistenter är fortfarande relativt outforskat, varav mallbaserad (template-based) generering fortfarande är det vanligaste sättet av

textgenerering i talassistenter [15]. Problemet med mallbaserad textgenerering är dock mängden resurser som krävs för att utveckla ett sådant system. Mallbaserad textgenerering är även svår att implementera för flera språk eller på en större skala.

En mer systematisk textgenerering utvecklades på 1990-talet, där tre olika steg används. Första steget innebär val av innehåll, där innehållet fås från dialoghanteraren. Sedan planläggs meningens struktur, det vill säga i vilken ordning orden kommer i, och till sist genereras den slutliga meningen. Problemet med systematisk textgenerering är dock även här mängden tid och data som behövs. För att minska på mängden tid och data som behövs, utvecklades en mer flexibel textgenerering som kan använda sig av ett korpus. Ett korpus är en stor mängd databehandlade texter, och fungerar som en databas [16]. Den korpusbaserade metoden för textgenerering blev användbart tack vare stordataperioden och gav möjligheten att generera text på basen av tidigare data samlat av talassistenten. Textgenerering med hjälp av korpusmetoden möjliggör även en mycket mer mänskligt och naturlig meningsstruktur. [15]

Tack vare framsteg gällande djupa neurala nätverk under de senaste årtiondet, har även textgenerering med hjälp av återkommande neurala nätverk (Recurrent Neural Networks, RNN) skapats [15]. Textgenerering med hjälp av RNN fungerar då djupa neurala nätverk används parallellt i talassistentens språkmodell. RNN kan därmed generera text genom att läsa indata från språkmodeller och strukturera meningar på basen av användarens indata och tidigare talhistorik. Textgenerering med hjälp av RNN även illustrerat i figur 2. Stjärnan i figuren indikerar det följande förutspådda ordet som formar texten samt varje *h* indikerar tidigare talhistorik. Pilarna till varje *h* visar hur varje ord i utdata sparas i talhistoriken. Textgenerering med hjälp av RNN skapar trots detta ännu vissa utmaningar. Textgenerering med RNN har framstått vara instabilt då det gäller långa textsekvenser, samt bearbetningen av långa indatasekvenser har visat sig vara väldigt tidskrävande [17].



Figur 2. Textgenerering med hjälp av RNN. Omarbetad och översatt från [[18]*, s. 47]*

3.6 Talsyntes (TTS)

Talsyntes, även kallad text-till-tal, är en process som kan generera människoliknande tal från inmatad text [19]. Processen kan implementeras på olika sätt, varav de traditionella sätten är genom att sammansätta vågformer (Waveform concatenation) eller genom att använda statistisk analys (Statistical parameter speech synthesis, SPSS). Talsyntes med hjälp av neurala nätverk har även implementerats under senare år för att ersätta de traditionella sätten.

Sammansättningen av vågformer fungerar genom att länka ihop fonemen som matas in som indata. Ett förinspelat korpus används för att jämföra hur fonemen skall sammanställas för att bilda fullständigt tal [19]. Tal som bildas genom att sammansätta vågformer är dock begränsat till talkorpuset, eftersom talsyntes genom sammansättning av vågformer endast kan producera tal som tidigare inspelats till det korpus som används [20].

Talsyntes genom att använda statistisk analys är dock inte begränsat till ett korpus, och löser därmed problemet. SPSS system tränas med hjälp av maskininlärning samt en stor mängd träningsdata och kan därmed producera tal.

Den vanligaste strukturen för SPSS innehåller en liknande akustisk modell som i komponenten automatisk taligenkänning, samt en ljudprocessor (vocoder) för att generera vågformer. Akustiska modellen i SPSS fungerar genom en kombination av en Gaussisk blandningsmodell och en Dold markovmodell (GMM-HMM) men i omvänd riktning jämfört med ASR. [20]. Talsyntes anses därmed vara den omvända formen av automatisk taligenkänning.

Neurala nätverk kan implementeras i talsyntes istället för att använda de traditionella sätten för att generera tal. Neurala nätverket WaveNet kan användas som ljudprocessor för att generera realistiskt och människolikt tal.

Neurala nätverk i talsyntes kan ersätta de mer traditionella sätten. Det neurala nätverket WaveNet är en av flera neurala nätverk som kan användas inom talsyntes för att omvandla text till tal [20]. WaveNet fungerar som en ljudprocessor i komponenten för att generera realistiskt och människolikt tal. Utvecklingen av flexibel talsyntes med hjälp av neurala nätverk är dock väldigt dyrt. På grund av detta är talsyntes med hjälp av neurala nätverk ännu ett aktivt forkningsområde.

4. Finns det talassistenter för åldringar?

Problemet med automatisk taligenkänning för åldringar är att taligenkänningssystem med djupa neurala nätverk oftast enbart fokuserat på träningsdata samlat av medelålders människor. För att ett taligenkänningssystem skall fungera bra för åldringar måste även träningsdata av åldringar använts. I en undersökning av The PaeLife project [21] samlades hundratals timmar med träningsdata gällande åldringar för att skapa en virtuell personlig assistent med namnet AALFred. AALFred funktionerar i samband med människo-datorinteraktion och kan därmed erbjuda åldringar möjligheten att kommunicera både genom tal, gester och genom användningen av en pekskärm. Denna avhandling kommer fokusera på kommunikation mellan åldringar och datorer med hjälp av automatisk taligenkänning.

4.1 Vilken nytta skulle det medföra?

Work in progress

4.2 Vilka utmaningar skulle man möta?

Work in progress

5. Hur kan man optimera en talassistent för åldringar?

I dagens läge finns det närmare 900 000 finländare som är över 70 år gamla [22]. Av dessa åldringar lider många av fysiska svårigheter som dålig syn, dålig hörsel samt svårgheter att använda händerna. Många äldre lider även av ensamhet, dåligt minne och depression. En lösning som skulle underlätta vardagen samt hjälpa mot ensamheten kunde därför vara en talassistent specifikt utvecklad för åldringar. Men hur skulle användargränssnitten för en talassistent kunna se ut?

Det mest naturliga sättet för åldringar att kommunicera är med hjälp av rösten, därav en digital assistent med ett talgränssnitt skulle fungera bäst. Ett talgränssnitt är ett användargränssnitt som främst eller enbart använder tal [?]. Talgränssnittet använder sig av en mikrofon för indata samt en högtalare för utdata. Utdatat framställs med hjälp av talsyntes, som är en teknik som försöker efterlikna en mänsklig röst.

Olika funktioner som en talassistent för åldringar kunde innehålla presenteras till följande. Talassistenten kan svara på olika frågor som ”vad är det för väder i morgon”, ”vad är klockan”, ”vilken tid kommer nyheterna på TV” och ”vem är finlands president”. Talassistenten kan även ringa efter hjälp om det hänt en olycka, till exempel om den äldre fallit och inte slipper upp, genom aktiveringsordet ”Hjälp”. Då hjälp ropas kontaktar talassistenten den äldres

kontaktperson som kan vara någon familjemedlem eller vårdpersonal om den äldre bor på åldringshem. Den äldres kontaktperson har åtgång till talassistentens inställningar och kan även lägga till påminnelser som till exempel påminner den äldre om morgonmedicinen och/eller kvällsmedicinen. Specifika påminnelser kan även läggas till, som till exempel påminner åldringen att äta, dricka eller röra på benen. Talassistenten kan även spela upp musik samt ljudböcker.

5.1 Vilken nytta skulle det medföra?

Work in progress

5.2 Vilka utmaningar skulle man möta?

Work in progress

6 Diskussion

Work in progress

7 Sammanfattning

Work in progress

Källförteckning

Work in progress

[1] “https://www.stat.fi/til/sutivi/2021/sutivi_2021_2021-11-30_tie_001_fi.html .”

[2] “https://www.stat.fi/til/sutivi/2013/sutivi_2013_2013-11-07_kat_001_fi.html ”.

[3] A. A. Librarian, I. Org |, and | Ieee-Sa, “Florida Institute of Technology Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home).” [Online]. Available: https://ieeexplore-ieee-org.portal.lib.fit.edu/document/8301638/citations?part=1

[4] D. Y. Zhang, J. Hueser, Y. Li, and S. Campbell, “Language-Agnostic and Language-Aware Multilingual Natural Language Understanding for Large-Scale Intelligent Voice Assistant Application,” Jan. 2022, pp. 1523–1532. doi: 10.1109/bigdata52589.2021.9671571.

[5] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, “Introduction,” *Robust Automatic Speech Recognition*, pp. 1–7, Jan. 2016, doi: 10.1016/B978-0-12-802398-3.00001-5.

[6] S. Rawat, P. Gupta, and P. Kumar, “Digital life assistant using automated speech recognition,” in *Proceedings of the International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity, CIPECH 2014*, Jan. 2014, pp. 43–47. doi: 10.1109/CIPECH.2014.7019075.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2016, vol. 2016-May, pp. 4960–4964. doi: 10.1109/ICASSP.2016.7472621.

[8] N. Desai, P. Dhameliya, and P. Desai, “International Journal of Emerging Technology and Advanced Engineering Feature Extraction and Classification Techniques for Speech Recognition: A Review,” *Certified Journal*, vol. 9001, no. 12, 2008, Accessed: Feb. 27, 2022. [Online]. Available: www.ijetae.com

[9] R. Carlson, “Datadrivna metoder för para-metersyntes-beskrivning av ett system och experiment med CART-analys Tor Sigvardson.”

[10] “https://svenska.se/tre/?sok=homofon&pz=1 .”

[11] E. Atwell, C. Souter, G. E. Churcher, and E. S. Atwell, “Dialogue Management Systems: a Survey and Overview Ensemble Morphosyntactic Analyser for Classical Arabic View project Islamic application of automatic question answering system View project SCHOOL OF COMPUTER STUDIES RESEARCH REPORT SERIES Dialogue Management Systems: a Survey and Overview,” 1997, doi: 10.13140/2.1.4252.1283.

[12] M. Mctear, “Dialogue Management The Role of Spoken Dialogue in User-Environment Interaction.”

[13] “https://smarterhomeguide.com/alexa-search-engine/ .”

[14] M. Riou, B. Jabaian, S. Huet, F. Lefèvre, and F. Lefèvre FABRICELEFEVRE, “Reinforcement adaptation of an attention-based neural natural language generator for spoken dialogue systems Reinforcement adaptation of an attention-based neural natural language generator for spoken dialogue systems Stéphane Huet,” vol. 10, no. 1, pp. 1–19, 2019, doi: 10.5087/dad.2019.101ï.

[15] T. H. Wen and S. Young, “Recurrent neural network language generation for spoken dialogue systems,” *Computer Speech and Language*, vol. 63, Sep. 2020, doi: 10.1016/j.csl.2019.06.008.

[16] “https://svenska.se/tre/?sok=korpus&pz=1.”

[17] A. Madotto, C.-S. Wu, and P. Fung, “Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems,” Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.08217

[18] J. D. Kelleher and S. Dobnik, “What is not where: the challenge of integrating spatial representations into deep learning architectures Learning language with robots View project Arabic Dialects NLP View project What is not where: the challenge of integrating spatial representations into deep learning architectures.” [Online]. Available: https://www.researchgate.net/publication/325385139

[19] J. Hu and A. Hamdulla, “Research on the Methods of Speech Synthesis Technology,” in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning, PRML 2021*, Jul. 2021, pp. 227–233. doi: 10.1109/PRML52754.2021.9520718.

[20] “https://wiki.aalto.fi/display/ITSP/Statistical+parametric+speech+synthesis .”

[21] A. Hämäläinen, A. Teixeira, N. Almeida, H. Meinedo, T. Fegyó, and M. S. Dias, “Multilingual Speech Recognition for the Elderly: The AALFred Personal Life Assistant,” *Procedia Computer Science*, vol. 67, pp. 283–292, 2015, doi: 10.1016/J.PROCS.2015.09.272.

[22] “https://www.stat.fi/til/vaerak/2019/vaerak_2019_2020-03-24_tie_001_fi.html .”