2023-10-27

Litteraturstudie med tillämpning av AI-baserade digitala tjänster

What is Murf?

Kurs: Informationssystem och data - C1ID1A
Termin: HT2023
Examinator: XX
Författare: XXXX

Sammanfattning

I en värld där artificiell intelligens blir allt mer utbredd vill vi undersöka dess påverkan på en traditionell tjänst - röstsyntes. Avsikten med denna rapport är att undersöka och analysera hur applikationen Murf använder AI för att förbättra röstsyntes samt utvärdera användbarheten i denna utvecklade produkt. Rapporten grundar sig i en litteraturstudie och sedan skapar vi ett projekt i Murf för att kunna utvärdera tjänsten. AI-tekniker som Murf använder och rapporten fördjupar sig i är naturlig språkhantering, artificiella neuronnät samt Generative Adversarial Networks.

INNEHÅLLSFÖRTECKNING

INLEDNING

Under de senaste åren har artificiell intelligens (AI) genomgått en anmärkningsvärd expansion, vilket har lett till en revolution inom många sektorer av vårt samhälle. Från medicinska framsteg, som hjälpmedel för att snabbt kunna identifiera cancer, till att förbättra vår vardag genom teknik som ansiktsigenkänning för upplåsning av mobila enheter, har AI visat sig vara en kraftfull resurs. Denna teknologiska utveckling har inte bara förbättrat effektiviteten och precisionen inom dessa områden, utan har också öppnat dörren för innovationer som tidigare var otänkbara. Ett sådant område är röstsyntes, där AI:s förmåga att efterlikna mänskligt tal som öppnar möjligheten till nya användningsområden. I denna rapport kommer vi att utforska de tekniska komponenterna bakom denna tillämpning av AI och dess betydelse i dagens digitala miljö.

Bakgrund

Ursprungligen var text-till-tal-system mekaniska och begränsade i sin förmåga att efterlikna mänskligt tal (Gold *et al.* 2011, s. 9-13). Med tiden, tack vare framsteg inom artificiella neuronnät och djupinlärning, har dessa system blivit alltmer sofistikerade och kan nu generera tal som är nästan omöjligt att skilja från en mänsklig röst.

Murf är en framstående aktör inom denna teknologiska utveckling. Det är en molnbaserad AI-tjänst som omvandlar text till tal och genererar människoliknande uttal. Användare har möjlighet att välja mellan en mängd olika språk, dialekter och tonlägen. Dessutom kan man anpassa den genererade rösten för att bättre efterlikna personer av specifikt kön och åldersgrupp. Det går även att lägga till bilder, musik och videoklipp till rösten.

De listar en stor bredd av användningsområden på sin hemsida. Där inkluderas bland annat uppläsning av böcker, röstskådespel till karaktärer i spel och animerad video samt röster till instruktioner och reklam. Medan dessa funktioner kan vara fördelaktiga för privatpersoner, framförs deras huvudsakliga nytta inom företagssektorn. Genom att använda Murf kan företag spara både tid och pengar, eftersom traditionell röstinspelning kan vara både kostsam och tidskrävande.

Syfte

Målet med denna rapport är att fördjupa förståelsen för den webbaserade tjänsten Murf, som bygger på artificiell intelligens, och därigenom ge en omfattande översikt över dess kapacitet, tekniska grund och potential inom den digitala marknadsföringsvärlden.

Frågeställning

Vi utgår från följande frågor:

1. Vilka underliggande AI-tekniker används av Murf?
2. Vilka praktiska tillämpningar finns för Murf?

Avgränsning

Murf erbjuder i dagsläget inte stöd för svenska, vilket har lett till att vårt valda scenario och den genomförda litteraturstudien baseras på engelska som språk. Denna avgränsning är viktig att notera eftersom den kan påverka generaliserbarheten av våra resultat, särskilt i en svensk kontext. Dessutom kan detta påverka utfallet av vårt Turing-test, eftersom testgruppen huvudsakligen består av personer med svenska som modersmål. Det är därför viktigt att närma sig våra slutsatser med förståelse för denna språkliga begränsning.

Metod

Denna rapport använder en kombinerad metodik som inkluderar praktiska användande av Murf och en litteraturstudie för att identifiera och förstå de underliggande AI-teknikerna.

GENOMFÖRANDE

Litteraturstudie

För att identifiera och förstå de AI-tekniker som Murf använder, inledde vi med att granska informationen tillgänglig på Murfs officiella webbplats. Vidare kompletterade vi denna information med insikter från vår kurslitteratur av Håkansson & Hartung (2020), där vi fick en överblick av delområden inom AI.

För att validera och fördjupa vår förståelse av dessa tekniker, genomförde vi en systematisk sökning av vetenskapliga artiklar via databasen Primo. Vi begränsade vår sökning till "Peer Reviewed"-artiklar och använde sökord som var relaterade till AI-teknikerna som identifierades i Murf. Varje källas relevans och trovärdighet verifierades genom att granska författarnas bakgrund, publiceringsplats och antal citeringar. Ett undantag från vår uppsatta regel om Peer Reviewed finns, då Goodfellow, I. et al. (2020) inte är Peer Reviewed. Vi valde att ändå använda den som källa då den

besvarar vår frågeställning. Författaren är framstående inom området den beskriver samt att artikeln blivit citerad över 20000 gånger.

Praktiskt utförande med Murf

För att konkret utforska Murfs förmågor och användbarhet valde vi att skapa en ljudbaserad reklam avsedd för plattformar som Spotify eller radio. Valet av reklam som uppgift baserades på vår uppfattning att marknadsföring är ett fördelaktigt användningsområde för tjänster som Murf, särskilt i ett snabbt föränderligt samhälle där effektiv och snabb reklamproduktion är avgörande.

Processen var som följer:

1. Kontoskapande: Vi skapade ett konto på Murf genom att logga in via Google.
2. Projektval: Efter inloggning kommer man till en startsida där man kan välja olika alternativ på vilken typ av projekt man vill ha, eller om man vill börja med ett tomt projekt. Vi valde att skapa ett nytt projekt med typen “Audio Ad”.
3. Textbearbetning: Vi infogade ett förberett manus från Murfs hemsida på engelska, eftersom svenska inte stöds av tjänsten. Vi väljer “Split by paragraph” som delar och gör paus mellan varje stycke automatiskt, men man kan även lägga in pauser manuellt. För manus, se Bilaga 1.
4. Röst- och musikval: Vi valde engelska (US) som språk, och ställde in att rösten skulle vara från en medelålders man. Av de alternativ som finns väljer vi rösten "Finn" för uppläsningen och lade till ljudspår med titeln “New Energy” från Murfs bibliotek.
5. Export: Slutprodukten, en ljudfil, exporterades från Murf.ai. Denna fil var strax under en minut lång och innehöll både röst och musik.

LITTERATURSTUDIE

AI-tekniker i Murf

De tekniker som används i Murf är naturlig språkbearbetning (*Natural Language Processing, NLP*), artificiella neuronnät (*Artificial Neural Networks, ANN*) samt Generative Adversarial Networks (GAN).

I boken av Håkansson & Lee Hartung (2020, s. 22-24) ges en översiktlig presentation över de olika grenarna inom AI. En betydande gren inom AI är NLP, som fokuserar på interaktion och kommunikation genom text och tal. Denna teknik utgör en central del av Murf.

Murf beskriver på sin hemsida tekniken för neural text-to-speech vilket gör att vi med hög säkerhet kan anta att de använder ANN i någon utsträckning. Håkansson och Lee (2020, s. 25) skriver om ANN som del av maskininlärning. I boken (2020, s. 248) beskrivs NN som ett bra val när källdatan uppvisar komplexitet, vilket mänskligt tal anses att ha. Vidare beskriver de även djupa neurala nätverk, och speciellt Feedforward Neural Networks (FFNN), som extra lämpliga för att analysera text och ljud.

GAN är en generativ teknik som lämpar sig väl för ett program som Murf, och tekniken omnämns på Murfs hemsida. Det är en AI-teknik baserad på spelteori och djupa neurala nätverk, och är specialiserad på att generera data som är realistisk.

Naturlig språkbearbetning

I en artikel skriven av Julia Hirschberg och Chistopher D. Manning (2015) diskuteras NLP och dess användande i verkliga scenarion. De framför att syftet med artikeln är att diskutera framgångar och utmaningar med implementering av NLP i olika sammanhang.

NLP använder sig av datavetenskapliga tekniker för att förstå och skapa mänskligt språkliga verk. Tidigare metoder fokuserade på att automatisera analysen av språklig struktur och utveckla grundläggande teknologier som maskinöversättning, taligenkänning (*eng. speech recognition*) och röstsyntes (*eng. speech synthesis*). Forskare tillämpar dessa verktyg i talande dialogsystem, tal-till-tal-översättning och utvinning av data från sociala medier (Hirschberg & Manning, s. 261-262).

Artikeln (Hirschberg & Manning, 2015, s. 262) belyser även de utmaningar som finns inom maskinöversättning, särskilt behovet av att djupt förstå mänskligt språk för att uppnå effektiv och korrekt översättning. En betydande begränsning av NLP är att de flesta resurser och system endast finns tillgängliga för språk med höga resurser som engelska, franska, spanska, tyska och kinesiska. Många språk med låga resurser saknar sådana resurser.

Slutsatser som dras i artikeln (Hirschberg & Manning, 2015, s.265-266) är att stora förbättringar inom speech recognition har skett. Som följd har det blivit vanligt att prata med sin telefon (t.ex. Apple Siri, Google Assistant och Microsoft Cortana), webbsökmotorer blir allt bättre på att förstå komplexa frågor, och maskinöversättning kan ge en grundläggande förståelse av material på ett annat språk, även om den ännu inte kan ge översättningar av mänsklig kvalitet.

På kort sikt tror författarna att en kombination av mer data och ytterligare framsteg inom djupinlärning kommer leda till ytterligare framsteg inom NLP. Det räcker dock inte med tekniska framsteg utan kommer även krävas framsteg inom lingvistik, läran om mänskliga språk (Hirschberg & Manning, 2015, s. 265-266).

Artificiella neuronnät

Reddy och Rao (2013) framhåller i artikeln vikten av en väl fungerande intonationsmodell för att generera syntetiskt tal som låter trovärdigt och är behagligt att lyssna på. Prosodi är en gren inom fonetiken som studerar rytm och melodi i tal och intonation är i sin tur ett begrepp inom prosodi som beskriver tonhöjdsvariation och betoning. De vill därför undersöka hur en intonationsmodell fördelad på två steg, där de använder FFNN för att förutsäga intonationsmönster för en följd stavelser fungerar i jämförelse med andra tekniker (Reddy & Rao 2013, s.1106).

Reddy och Rao (2013, s. 1106) beskriver NN som extra lämplig för TTS då de är kända för sin förmåga att upptäcka mönster i det material den tränats på. NN har också generaliseringsförmåga för att kunna förutsäga intonationsmönster någorlunda bra för de mönster som inte fanns under inlärningsfasen.

Många tekniker har utvecklats, men två tillvägagångssätt för modellering av intonation har varit extra framträdande under de senaste 20 åren, tonsekvensmetoden som följer traditionell fonologisk beskrivning av intonation och superpositionsmetoden. Reddy och Rao (2013, s. 1106-1107) beskriver dessa modeller, deras styrkor och brister tillsammans med en rad ytterligare modeller för intonation, både med och utan neurala nätverk som teknik. De ger även exempel inom specifika språk.

Reddy och Rao (2013, s. 1124) slår fast att en intonationsmodell med två steg och som använder neurala nätverk visade goda resultat. Modellen de beskrivit i studien visade på en hög prediktionsnoggrannhet och när man låtit personer lyssna på resultat från modellen visar det på hög kvalitet, högre än andra modeller som beskrivs i studien.

Generative Adversarial Networks

Goodfellow, I. et al. (2020, s. 139) inleder artikeln med att i korthet beskriva en rad olika generativa modeller, tillsammans med styrkor och svagheter. Generativa modeller har som syfte att skapa något, såsom bilder, språk eller ljud. De bedömer att generativa

modeller baserade på övervakad inlärning (*eng. supervised learning*) ofta har en precision som är bättre än mänsklig efter träning, men att de under träningsfasen kräver både stora mängder data att träna på och mycket mänsklig övervakning. Övervakade inlärningsmodeller i stort har samma mål, att kunna koppla källdata till rätt måldata. Modeller baserade på oövervakad inlärning (*eng. unsupervised learning*) är istället mindre konkret då modellerna kan ha många olika mål. För generativa modeller kan man t.ex. ta träningsexempel från en samling med okänd fördelning och låta modellen ta fram en algoritm som uppskattar fördelningen så korrekt som möjligt.

GAN är baserad på spelteori, innehållande två "spelare". Den ena är en generativ modell som beskrivs i korthet ovan, som brukar kallas för generator. Generatorn har som uppgift att generera en fördelning, som en ljudfil om vi använder Murf som exempel. Den andra spelaren är en urskiljare *(eng. discriminator*) vars uppgift är att kontrollera generatorns ljudfil för att se om det är riktig data från träningsunderlaget eller skapad av generatorn. Om urskiljaren lyckas upptäcka att datan den fått är förfalskad förlorar generatorn "spelet" och tvingas se över sin taktik (eg algoritm) för att skapa bättre förfalskningar. Misslyckas urskiljaren har den istället förlorat och tvingas se över sin taktik (algoritm). Goodfellow, I. et al. (2020, s. 140-141) använder förfalskare av sedlar och polisen som exempel för att ytterligare förtydliga processen. Då de två spelarna konstant ger återkoppling till varandra krävs ingen mänsklig övervakning.

Målet för maskininlärningsalgoritmer är att uppnå ett lokalt Nash equilibrium. Goodfellow, I. et al. (2020, s. 142) beskriver det som en punkt där spelarna känner till varandras strategier och ingen har något att vinna på att ändra sin egen strategi.

Goodfellow, I. et al. (2020, s. 144) drar slutsatsen att GAN är väldigt framgångsrikt när det gäller att generera realistisk data men det är ännu svårt att träna dem. För att göra modellen mer pålitlig kommer det behöva utvecklas mer för att kunna nå Nash Equlibrium snabbt och konsekvent.

ANALYS & DISKUSSION

Under vårt praktiska genomförande av Murf kan vi se var vardera AI-teknik appliceras och påverkar slutprodukten. När vi valt projekt och lägger in texten vi vill syntetisera, eg. “manus”, ser vi hur NLP används för att dela upp ordföljden och kategorisera orden efter plats i meningen och ordtyp.

När vi därefter ställt in alla val om hur rösten ska låta och väljer att generera ljudfilen så tar ANN och GAN över för att applicera melodi, tonläge och betoningar utefter informationen från NLP-tekniken. Detta ger ett slutresultat som är långt mer trovärdigt än traditionell röstsyntes.

Vår reflektion är att slutprodukten är tillräckligt trovärdig för att vara behaglig att lyssna på, även under längre perioder. I dagsläget är den däremot inte likvärdig en mänsklig röst och man kan i vissa fall urskilja att det handlar om röstsyntes. Den är känslig för stavfel, gemener och versaler samt styckesindelning. Det krävde en del arbete att få rösten att betona ord och ta korta uppehåll på ett trovärdigt sätt, mer än några få minuter som sägs på hemsidan. En annan nackdel var att det var långsamt att processa ljudfilen. Däremot var användargränssnittet användarvänligt och intuitivt så att göra detta var inte svårt utan tog bara lite tid. Det var även enkelt att lägga till musik eller video.

Vår utgångspunkt var att skapa ett ljudspår för marknadsföring. Efter att ha genererat en ljudfil enligt den är vår slutsats att ljudfilen tydligt kommunicerar det manus vi lagt in utan fel i uttal eller betoning. Vår bedömning är att resultatet är tillräckligt trovärdigt för att agera ljudspår i till exempel radioreklam, men sannolikt inte så trovärdig att den förväxlas med mänskligt tal.

Vi tror att tjänsten kan vara lämplig vid uppläsning av ljudböcker, instruktioner samt röst till animerad film. En annan intressant applicering av tjänsten hade varit som hjälpmedel för personer som inte kan prata, men i dagsläget ser vi att genereringsprocessen är för långsam för att uppfylla det syftet.

SLUTSATSER

De underliggande AI-tekniker vi har identifierat i Murf är naturlig språkbearbetning (NLP), artificiella neuronnät (ANN) samt Generative Adversarial Networks (GAN). Appliceringen av AI i text-till-tal har resulterat i en hörbar förbättring i trovärdighet, och vi kan se flera potentiella användningsområden, såsom reklam och uppläsning av böcker och instruktioner.

REFERENSER

Gold, B., Morgan, N. and Ellis, D. (2011). *Speech and audio signal processing*. John Wiley & Sons (s. 9–13)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020) Generative adversarial networks. *Communications of the ACM*, 63(11), s. 139-144. doi: https://doi.org/10.1145/3422622

help.murf.ai. (n.d.). What is Murf? https://help.murf.ai/what-is-murf. [2023-10-03]

Hirschberg, J. & Manning, C. D. (2015) Advances in natural language processing. *Science (American Association for the Advancement of Science),* 349 (6245), s. 261–266. https://www.jstor.org/stable/24748572

Håkansson, A. & Hartung, R. L. (2020) *Artificial intelligence : concepts, areas, techniques and applications*. Studentlitteratur. (s. 22-25)

Murf.ai (2023). Generative AI : All you need to know https://murf.ai/resources/generative-ai/. [2023-10-13]

Murf.ai. (n.d.). Text to Speech Online: Generate realistic TTS voiceovers https://murf.ai/text-to-speech. [2023-10-13]

Reddy, V.R. and Rao, K.S. (2013). Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis. *Computer Speech & Language,* 27(5), pp.1105–1126. doi:https://doi.org/10.1016/j.csl.2013.02.003.

Bilaga 1

Manus:

You can say goodbye to the traditional and tedious voiceover production process with Murf Studio.

Murf Studio is a cloud-based realistic text-to-speech platform that allows you to generate lifelike AI-based voiceovers for multiple use cases like Youtube videos, Podcasts, Marketing & Advertising, Audiobooks, Games, Product & Explainer videos, and more.

Our AI-based platform simplifies the entire process and saves you time and money. With over 120 voices in 20 languages, you can easily convert your script into high-quality audio within minutes.