Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Fredrik Bache Ruud
Martin Nordli Almenningen

Norsk tekst-til-tale med Tacotron 2 og WaveGlow

Bacheloroppgave i ingeniørfag, data
Veileder: Tomas Holt
Medveileder: Anna Kim & Knut Inge Hvidsten
Mai 2024

NTNU
Kunnskap for en bedre verden

pexip

Fredrik Bache Ruud
Martin Nordli Almenningen

Norsk tekst-til-tale med Tacotron 2 og WaveGlow

Bacheloroppgave i ingeniørfag, data
Veileder: Tomas Holt
Medveileder: Anna Kim & Knut Inge Hvidsten
Mai 2024

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

NTNU

Kunnskap for en bedre verden

Sammendrag

Denne oppgaven omhandler utviklingen av en norsk Text-til-tale (TTS)-modell ved hjelp av maskinlæringsmodellene Tacotron 2 og WaveGlow. Disse modellene er designet for å generere TTS-modeller, men det finnes begrenset dokumentasjon om deres anvendelse på andre språk. Oppgaven vil demonstrere hvordan det er mulig å gjennomføre dette for mindre språk som norsk, som også inkluderer unike bokstaver.

Bruken av Tacotron 2 og WaveGlow for å utvikle TTS-modeller på den måten som beskrevet i denne oppgaven, vil teoretisk sett ikke begrense hvilke språk som kan anvendes. Det eneste kravet er et tilstrekkelig stort datasett for trening, og eventuelt en oppdatering av alfabetet for å inkludere nye bokstaver. Tidligere studier har demonstrert utviklingen av modeller for blant annet sanskrit, men også for andre språk. Utover WaveGlow kan man også se på eksempler som Googles WaveNet-modell, som har blitt brukt til å utvikle TTS-modeller på flere titalls forskjellige språk, basert på de samme prinsippene.

Oppgaven ble valgt på grunn av TTS-teknologiens økende betydning i hverdagen. Muligheten til å få tekst opplest av en digital enhet gir utallige muligheter for økt tilgjengelighet av data og nye anvendelser av digitale medier. Oppdragsgiveren, Pexip, har som mål å integrere denne teknologien i sin programvare for å hjelpe brukere med å høre blant annet oppsummeringer av møtetranskripter eller mottatte meldinger.

Arbeidet som er utført demonstrerer at det er mulig å utvikle en god TTS-modell med Tacotron 2 og WaveGlow på mindre språk, som norsk. Ved å fremheve muligheten for å lage effektive TTS-modeller med relativt begrensede datasett for mindre språk, kan man bidra til å forbedre digitale verktøy ytterligere. Selv om modellen ikke er perfekt, representerer den et skritt nærmere maskiner som kan snakke like flytende som mennesker.

Abstract

In this thesis the result of making a Norwegian Text-to-Speech (TTS) model will be presented. The model has been created using the machine learning models Tacotron 2 and WaveGlow. These are machine learning models that are designed to make TTS models. Documentation about how to create models in other languages than English is limited. The goal of this thesis is to shine some light on how this is possible even for small languages like Norwegian, that also has its own letters that must be considered.

Using Tacotron 2 and WaveGlow to create a TTS model, as described in this report, is theoretically not restricted to the original language. The only requirement is that you have a dataset that is big enough for training, and if necessary, you must update the alphabet to include all required letters. Other than that, it is the algorithms task to figure out what is important for the selected language, and relatively few things must change. From other examples, models have been created in other languages already. There is a thesis that was written for Sanskrit, and other examples are also available. If you look outside WaveGlow, you can find examles from Googles own WaveNet model. That model is used to create models in tens of languages and learns using the same principles.

The reason this thesis was chosen is because of the significance TTS models have on our everyday life. The possibility to have any text being read out to us from a digital device opens many new possibilities to create new accessible ways to understand data. The company that requested the thesis to be done, Pexip, wants it to be able to read out summarized meeting transcripts, or received messages.

The work done illuminates the possibilities of making a good TTS model using Tacotron 2 and WaveGlow in languages such as Norwegian. By showing off the possibilities to create good TTS models with a relatively limited dataset in smaller languages, opens the door for more good solutions for making digital tools more available and accessible. The model is in no way perfect, but a step closer to making machines as good as humans to speak to us.

Forord

Oppgaven ble valgt da begge var interessert i temaet og ønsket å jobbe videre med maskinlæringskunnskapene sine. Tidlig arbeid gitt ut på å planlegge prosessen, og sette opp samarbeidsavtaler. Deretter hadde gruppen en opplesningsrunde der vi leste oss opp på oppgaven og verktøyene som det var nødvendig å lese om. Her samlet gruppen mye informasjon som gjorde det lettere å ta valg senere i oppgaven, og ga ett godt grunnlag for å jobbe med den. Deretter begynte prosessen som var å lage og utvikle datasett, og finne ut av hvordan man trener modellen. Dette tok lengre tid enn først forventet, men arbeidet her sikret det gode resultatet gruppen har fått. Før det siste steget i oppgaven var å dokumentere, og rapportere.

Etter et langt semester er det nå på tide å presentere prosessen og resultatene. Takk til Anna Kim, og Knut Inge Hvidsten fra Pexip, dere har vært til stor hjelp, og har gitt oss gode tilbakemeldinger og retning for denne oppgaven. Takk også til Tomas Holt for hjelp med skriving og assistanse til tilgang til NTNUs resurser. Uten deres hjelp hadde ikke resultatene vært i nærheten av det de er nå.

Fredrik Bache Ruud
21.05.2024

Martin Almeninngen
21.05.2024

Oppgavebeskrivelse

Problemstilling

Det finnes for øyeblikket mange variasjoner av det engelske språket for Text-til-tale (TTS) syntese tilgjengelig kommersielt. Utvalget er betydelig mindre for andre språk som tysk, fransk eller spansk. Dette gjelder i enda større grad for mindre språk som norsk, uten å ta hensyn til ulike dialekter.

Kort beskrivelse av oppgaveforslag

I en Pexip videokonferansesamtale ønsker vi å tilby en tekst-til-tale-funksjon for ikke-engelsktalende. En av bruksområdene er når du har opprettet et møte-transkript, og du vil at systemet skal lese opp sammendraget av samtalen, generert av kunstig intelligens.

Utfyllende kommentarer til hva oppgaven gjelder

Mål 1: Utvikle en ikke-engelsk akustisk modell (for eksempel i mel-spektrum eller lignende) for studentenes valgte språk som kan fungere i samarbeid med Nvidia WaveGlow for talesyntese. Den resulterende modellen skal fungere både i offline- og strømmemodus og kan distribueres på en Nvidia Riva-server.

Bonusmål: Integrer løsningen i Pexip Infinity.

Innhold

Figurer

Tabeller

Ordliste

arpabet ARPAbet er et fonetisk alfabet utviklet av Advanced Research Projects Agency (ARPA) for å representere engelske lyder i tekstform. ARPAbet består av en rekke symboler, der hvert symbol representerer en bestemt lyd (fonem) i engelsk tale. For eksempel representerer ”AH” lyden i ordet ”cot” og ”EH” lyden i ordet ”bet”. 16

Batch størrelse refererer til antall treningseksempler som brukes per iterasjon før modellen oppdateres.. 19

epoke En epoke er en enkel gjennomkjøring av hele treningdatasettet under treningen av en modell.. xiii, 19–21

Læringsrate er en parameter som brukes i maskinlæring for å bestemme hvor mye en modell sine vekter skal justeres i respons til feil hver gang modellen lærer. en høy læringsrate kan gi en modell som lærer raskt, men som kan bomme på den ”beste” løsningen. En lav læringsrate betyr at modellen tar mindre steg, og kan bruke mye ekstra tid på å forbedre seg. Å balansere dette godt er viktig for å få en god modell.. 19

vektene I maskinlæring er vekter (weights) numeriske verdier som justeres under treningen av en modell for å minimere feilen mellom modellens prediksjoner og de faktiske verdiene. Vektene bestemmer hvor mye påvirkning hver input-funksjon har på modellens utgang. . 16

Vekttap er en funksjon som passer på at en modell ikke overtrenes. Dette gjøres for at en modell ikke skal feste seg for mye til en enkelt egenskap til data og passer på at modellen lærer riktig.. 19

Akronymer

AMP	Automatic Mixed Precision.
CPU	Hovedprosessor.
CUDA	Compute Unified Device Architecture.
cuDNN	CUDA Deep Neural Network.
DSR	design science research.
GPU	Grafikkprosessor.
LPC	lineær prediktiv koding.
LTSM	Long Short Term Memory.
MCD	Mel-Cepstral Distortion.
Seq2Seq	Sekvens-til-sekvens.
TTS	Text-til-tale.

1. Introduksjon og relevans

Teknologien for Text-til-tale (TTS) har utviklet seg raskt de siste årene, og dens potensial for å forbedre kommunikasjon og tilgjengelighet i digitale applikasjoner er betydelig. Denne oppgaven fokuserer på å utvikle en TTS-modell på norsk til bruk i Pexips applikasjoner, med mål om å gjøre deres plattform mer inkluderende og tilgjengelig for alle brukere. Pexip, som leverer videokonferanseløsninger, kan med en integrert TTS-modell tilby en løsning som gjør det lettere for personer med lese- eller talevansker å delta fullt ut i møter og samtaler. Ved å tilby et verktøy som kan konvertere tekst til naturlig lydende tale, kan man åpne nye dører for effektiv kommunikasjon og inkludering.

For å gjennomføre prosjektet var det flere problemstillinger som dukket opp. Hvordan kan man skaffe seg et stort nok datasett som passer til trening? Hvordan trener man en TTS modell? Hvordan kan man vurdere resultatene og bekrefte kvaliteten på det ferdige resultatet? Svarene vil bli besvart gjennom denne rapporten. Den begynner med utdypning av viktige prinsipper, og teknologier er viktig. Før metoden anvendt i perioden blir lagt frem. Resultatene bli presentert og diskutert, før oppgaven blir konkludert.

2. Teori og relevant litteratur

2.1 Syntetisk tale

Syntetisk tale, eller talegenerering, har utviklet seg betydelig siden starten på 1930-tallet med VODER, og har gått gjennom teknologier som lineær prediktiv koding (LPC). Moderne teknikker for talegenerering har avansert langt utover disse opprinnelige metodene, og benytter dyp læring for å skape stadig mer naturlig og menneskelig tale.

Det har blitt utforsket flere tilnærminger innenfor feltet talegenerering over årene. Formantsyntese, konkatenativ syntese og parametrisk syntese har alle spilt sentrale roller i utviklingen av tidligere systemer. Imidlertid har innføringen av dyp læring markert en ny æra preget av modeller basert på nevrale nettverk. Blant disse har sekvens-til-sekvens-modeller med oppmerksomhetsmekanismer, som Long Short Term Memory (LTSM) nettverk og Transformers, vist seg å være spesielt effektive.

Utviklingen av Tacotron 2 og NVIDIAs WaveGlow markerer et vesentlig fremskritt innen talegenerering. Tacotron 2, som opererer på et ende-til-ende-prinsipp, konverterer tekst til et mel-spektrogram ved hjelp av en sekvens-til-sekvens-modell med oppmerksomhetsmekanisme. Dette gjør at den kan fange opp taleens nyanser mer presist enn tidligere metoder. WaveGlow kompletterer Tacotron 2 ved å transformere mel-spektrogrammet til hørbar tale, og bruker et strømbasert generativt nettverk for å skape lyd av høy kvalitet.

Disse teknologiene byr på flere fordeler, som lavere forsinkelse, tilpasningsdyktighet i talestiler og toner, samt evnen til å skape mer uttrykksfulle taleformer. Likevel møter de visse utfordringer. Behovet for omfattende treningsdata, vanskeligheter med å modellere spesifikke dialekter eller emosjonelle nyanser, og utfordringen med å produsere høykvalitets tale i sanntid er noen av de mest betydelige hindrene.

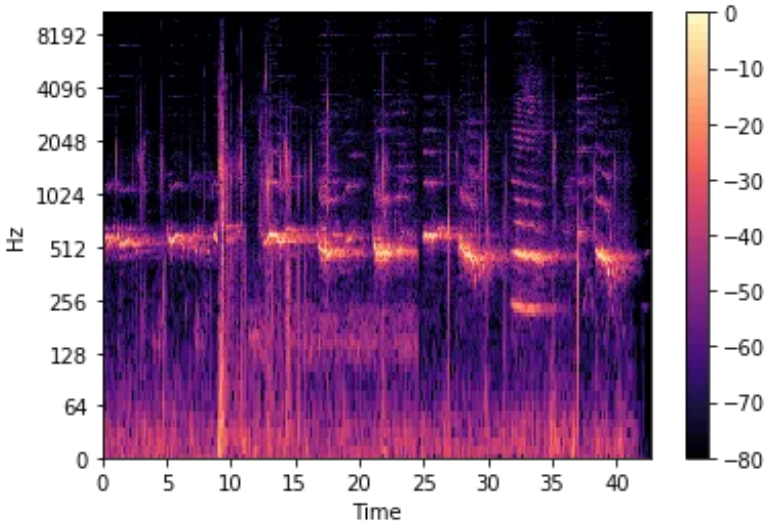
Syntetisk tale har et bredt spekter av bruksområder, fra drift av virtuelle assistenter og interaktive roboter til forbedring av pedagogiske verktøy og underholdningsteknologier. Ettersom disse systemene stadig forbedres, forventes det at deres innvirkning på hverdagslivet og ulike sektorer

vil øke, noe som understreker viktigheten av kontinuerlig forskning og utvikling i dette feltet.

2.1.1 Mel-spektrogram

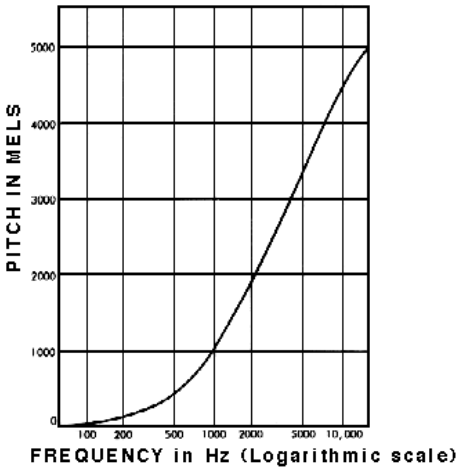
Et Mel-spektrogram er en måte å representere lyd, som en graf. Den kombinerer konseptene spektrogrammer med Mel skalaen for å gjøre det enda mer tydelig hvordan lydklipp faktisk høres ut for mennesker. For å forstå dette er det lurt med en forklaring av de forskjellige konseptene.

Et spektrogram er en visuell fremstilling av et spektrum av frekvenser av et lydsignal over tid. Den er typisk representert som en todimensjonal graf hvor tid er på x-aksen og frekvens er på y-aksen, med fargeintensitet som en indikator på styrken av frekvensene gitt i desibel.



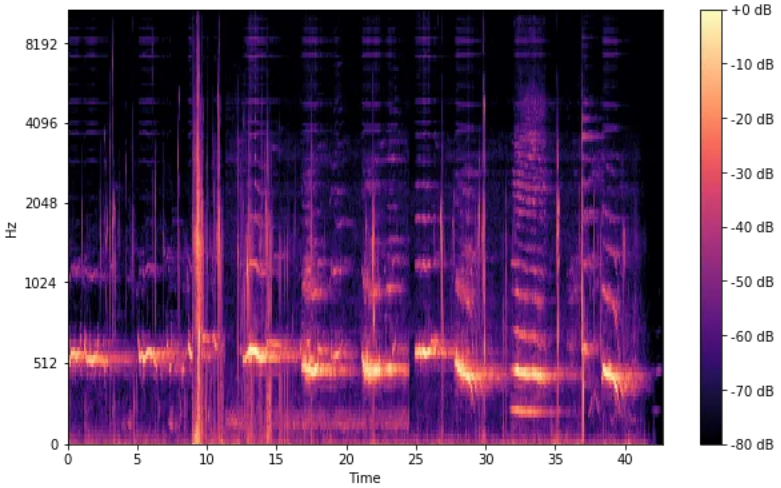
Figur 2.1: Spektrogram til et lydklipp. (Gartzman, 2019)

Mel skalaen er en tilpasset skala som skal dømme tonehøyde basert på menneskers respons på avstand mellom dem. Mennesker hører ikke forskjellen linjert. Forskjellen mellom 500Hz og 1000Hz er mye større enn forskjellen mellom 7500Hz og 8000Hz. Å forskyve grafen etter mel skalaen, hjelper med å fange opp mer meningsfylt data basert på det mennesker hører.



Figur 2.2: Mel skalaen (Doshi, 2021)

Siden et Mel-spektrogram er en kombinasjon av disse to konseptene, er dette den beste måten å visualisere hva mennesker faktisk hører. Det at skalaen er formet og designet etter menneskers hørsel hjelper fange opp flere viktige nyanser i språk, og er derfor et viktig verktøy i å reprodusere lyd.



Figur 2.3: Mel-spektrogram til samme lydklipp som vist i figur 2.1 (Gartzman, 2019)

2.2 Maskinlæring i TTS

Maskinlæring sin hovedoppgave går stort sett ut på å forutse eller klassifisere noe. Den gjør dette ved å lære gradevis over mange steg. Dette

er på samme måte som mennesker utvikler seg og lærer å kjenne igjen mønster. (IBM, 2024) Maskinlæringsmodeller lærer hovedsakelig basert på tre ulike metoder. Veiledet læring, Ikke-veiledet læring og forsterket læring. (Tidemann og Elster, 2023)

Veiledet læring går ut på at en maskinlæringsmodell lærer ved å sammenlikne resultater med eksisterende data. Den krever ofte store datasett og trenger ofte å være strukturert på en måte som er strukturert og lett å lese for en maskin. Ikke-veiledet læring fungerer annerledes. Den tar store umerkede datasett og prøver å finne mønster i datasett, og samler det i grupper. Forsterket læring den siste grenen innen maskinlæring. Denne modellen er ganske lik veiledet læring, men bruker ikke eksisterende data, men heller mål og retninger. Når en modell gjennomfører steg vil modellen utvikle seg videre på de beste resultatene, for å bli enda bedre.

2.2.1 Dyp læring og nevrale nettverk

Dyp læring kan fort bli blandet sammen med maskinlæring, dette stemmer ikke helt. Dyp læring er et undersett av nevrale nettverk som igjen er en underkategori av maskinlæring. Forskjellen i hvordan de lærer kan sammenliknes med veiledet læring, men den trenger ikke nødvendigvis å ha like sterke regler for struktur i datasett. Dyp læring kan løse mye av strukturen selv og gjør det mindre nødvendig for menneskelig inntreraksjon. Dette gjør at det er mindre muligheter for at data blir merket feil, og åpner opp for mye større datasett.

Hvis man setter denne teorien i kontekst med bruk i Tacotron, kan man se at selv om man har data som er koplet til taledata, ser man at ord og lyder er opp til modellen å finne ut av og kategorisere. Dataen er her semistrukturert, men har som oppgave å finne ut av hvordan separate ord, og uttalelser skal høres ut. Man slipper derifor å bruke mye tid på å legge inn og uttale alle ord og variasjoner av uttalelser, men lar maskinen lære seg hvordan uttalelser bør høres ut.

2.2.2 Konvolusjonelle nevrale nettverk

Et konvolusjonelt nevralt nettverk er en type dyp læringsalgoritme (Intel, 2020). De er ofte brukt til å analysere og lære visuelle trekk fra store mengder data. Konvolusjonelle nevrale nettverk fungerer ved å ta inn og prosessere data i et rutenett, før den trekker ut viktige små detaljer for klassifikasjon og deteksjon. De består ofte av tre typer lag; et konvolusjonell lag, et samlingslag, og et fult tilkoplet lag.

2.2.3 Long Short Term Memory

Long Short Term Memory (LTSM) er et type nevralt nettverk som kan beholde langsiktige avhengigheter i sekvensiell data (Banoula, 2023). LTSM kan analysere og prosessere som tekst og tale. De kan selektivt beholde eller fjerne innformasjon som gjør at de i mindre grad mister informasjon over tid som andre nevrale nettverk kan.

2.3 Teknologier

2.3.1 Docker og Nvidia container toolkit

Docker er en åpen kildekode plattform, som legger tilrette for enkel programvareutvikling, og styring av applikasjoner gjennom containere. En container er en enhet med kode som pakker opp kode og alle avhengighetene den trenger, sånn at koden kan kjøre på alle slags enheter. En Docker container er et system som inneholder alt som trenges for å kjøre en applikasjon. (Docker, 2024)

Nvidia container toolkit er verktøy som gjør at containere kan kjøre med GPU-akselerasjon. Verktøyet inneholder en container-kjørings bibliotek som automatisk konfigurerer containere til å benytte Nvidia sine GPU-er (Nvidia, 2024b).

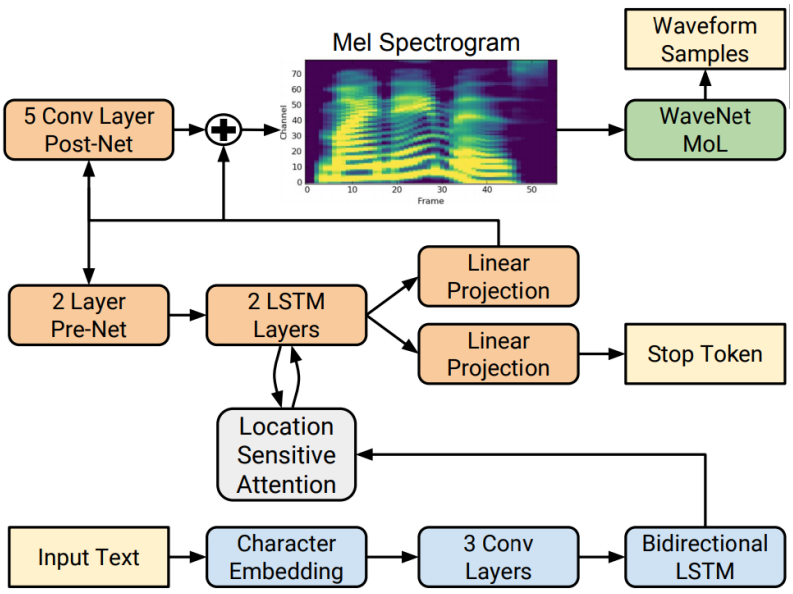
2.3.2 Pytorch

PyTorch er et åpen kildekode-rammeverk for maskinlæring som primært brukes for utvikling og trening av nevrale nettverk. Utviklet av Facebook's AI Research lab (FAIR), tilbyr PyTorch dynamisk grafkonstruksjon, noe som gjør det fleksibelt og intuitivt for forskere og utviklere. Rammeverket støtter automatisk differensiering, noe som er avgjørende for trening av dype nevrale nettverk, og det integreres godt med Python, som gir enkel debugging og en rik økosystem av verktøy og biblioteker.

2.3.3 Seq2Seq maskinlæringsmodell

Sekvens-til-sekvens (Seq2Seq) modeller er populære språkmodeller. Hovedideen er å spore en sekvens med inputer til en sekvens med outputer (Bayram Durna, 2024). Dette skjer i to lag, en encoder", og en decoder". Encoderen sin jobb er å lese og prosessere inputsekvenser, mens dekoderen sin jobb er å effektivt lære seg å generere den ønskede sekvensen, basert på konteksten av tidligere outputer.

2.3.4 Tacotron 2



Figur 2.4: Oversikt over tacotron 2 modellen fra PyTorch.

Tacotron 2 er en avansert TTS modell utviklet av Google, som består av to hovedkomponenter: en Sekvens-til-sekvens (Seq2Seq)-modell og en WaveNet-vocoder. Sekvens-til-sekvens (Seq2Seq)-modellen konverterer tekst til et spekter av mellombølgeformer, kalt mel-spektrogrammer. Deretter tar WaveNet-vocoderen dette mel-spektrogrammet og genererer høy kvalitet og naturlig tale. Tacotron 2 er kjent for sin evne til å produsere svært naturlig og uttrykksfull tale, noe som gjør den egnet for bruk i en rekke TTS-applikasjoner.

Fra Tekst til Tale: En Reise Gjennom Tacotron 2s Arkitektur

Innføringen starter med input-tekst som Tacotron 2 først behandler gjennom en karakter-embedding-mekanisme. Hvert tegn konverteres til et numerisk format, som er et viktig trinn for datamaskinens evne til å forstå og bearbeide språket.

Tekstsekvensene går deretter gjennom et 'pre-net', som består av to lag med ikke-lineære transformasjoner. Dette laget er essensielt for å redusere overtilpasning og fremme mer varierte utdata, og dermed gi modellen et mer robust grunnlag. Pre-nettet bidrar til å forhindre at modellen overtilpasser seg treningsdataene og sikrer bedre generalisering til nye, usette data.

Videre i prosessen benytter Tacotron 2 en encoder som inkluderer tre konvolusjonelle lag for å ekstrahere tidsuavhengige egenskaper, etterfulgt av et bidirectional LTSM-lag. Denne strukturen tillater en rik kontekstfangst ved å ta hensyn til informasjon både før og etter det aktuelle punktet i sekvensen.

Kjernen i modellens intelligens er den 'location sensitive attention'-mekanismen. Denne komponenten gjør det mulig for modellen å dynamisk fokusere på forskjellige deler av innteksten og bestemme hvilke deler av teksten som bør lydsettes med større vekt eller fokus.

Når riktig fokus er etablert, trer Tacotron 2s decoder i kraft med sine to LTSM-lag. Her skjer en sekvensiell produksjon av et mel-spektrogram, hvor hvert nytt stykke informasjon bygger på det foregående, og skaper en kontinuerlig strøm av lydinformasjon.

Deretter kommer 'post-net', som inneholder fem konvolusjonelle lag, inn for å forbedre det allerede genererte mel-spektrogrammet. Dette resulterer i et detaljrikt spektrogram klart for vokoderprosessen.

Parallelt med dette pågår en kontinuerlig forutsigelse av et 'stop token' fra et annet sett av projeksjoner. Denne funksjonen angir når modellen skal avslutte lydproduksjonen, noe som sikrer at talegenereringen blir naturlig i lengde og form.

Selv om den spesifikke vokoderen WaveGlow vil bli diskutert mer detaljert senere, er det verdt å nevne at uten en effektiv vokoder, ville mel-spektrogrammet forbli en visuell representasjon og ikke den naturlige tale vi hører. Tacotron 2s evne til å produsere et så detaljert og finjustert mel-spektrogram gjør det til et ideelt utgangspunkt for enhver avansert vokoder.

Denne gjennomgangen har dekonstruert Tacotron 2-modellens arkitektur, som vist i diagrammet, og avdekket de intrikate stegene modellen tar for å omdanne skrevet tekst til naturlig tale. Dette avsnittet er inspirert av Shen mfl., 2017

2.3.5 Automatic Mixed Precision (AMP) og mixed precision training

AMP (Automatic Mixed Precision) og mixed precision training er teknikker for å akselerere opplæring av nevrale nettverk ved å bruke både 16-bit (FP16) og 32-bit (FP32) flyttallsrepresentasjoner. Mixed precision training kombinerer disse for å redusere minneforbruk og øke ytelsen, siden FP16-beregninger er raskere og krever mindre minne. |AMP automatiserer den-

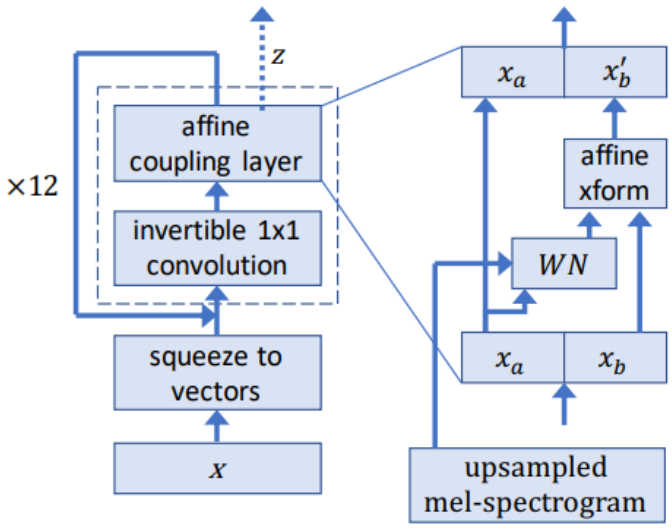
ne prosessen, noe som gjør det enkelt å bruke mixed precision i rammeverk som PyTorch og TensorFlow. Dette gir raskere treningstider, mulighet for større modeller og batch-størrelser, og optimal utnyttelse av moderne GPU-er som NVIDIA Tensor Cores.

2.3.6 Tensorboard

Tensorboard er et visualiseringsverktøy som hjelper med å overvåke treningsprosessen. Det gir innsikt til modellens opplæringsprosess ved å overvåke mål som læringsrate, tap, nøyaktighet, vektverdier og liknende. Dette hjelper utviklere med å feilsøke og optimalisere modellens effektivitet.

2.3.7 WaveGlow

WaveGlow er en maskinlæringsmodell som kan generere høykvalitets lydklipp fra Mel spektrogrammer. WaveGlow er implementert ved hjelp av ett enkelt nettverk, som gjør treningen enkel og stabil.



Figur 2.5: Figur over WaveGlow sitt nettverk (Prenger mfl., 2018)

WaveGlow: En detaljert gjennomgang av arkitekturen

WaveGlow er en modell som benyttes for å generere tale fra mel-spektrogrammer, og den kombinerer elementer fra Glow og WaveNet for å oppnå effektiv og naturlig lydsyntese. Diagrammet illustrerer de viktigste komponentene og prosessene i WaveGlow-arkitekturen.

Prosessen starter med input-vektoren x, som representerer det oppsam-

lede mel-spektrogrammet. Herfra går dataene gjennom flere transformasjonslag for å generere en naturlig og flytende lydsekvens.

Squeeze to Vectors

Første steg i prosessen er å 'klemme' input-vektoren x til en mindre representasjon. Dette gjøres for å redusere dimensjonaliteten og gjøre dataene mer håndterbare for de påfølgende lagene.

Invertible 1x1 Convolution

Etter å ha klemt vektoren, går dataene gjennom en inverterbar 1x1 konvolusjon. Dette laget tillater en ikke-lineær transformasjon av dataene, samtidig som det bevarer informasjonsflyten bakover gjennom modellen. Dette er essensielt for å oppnå høy kvalitet på den genererte lyden.

Affine Coupling Layer

Deretter går dataene gjennom et affine koblingslag, som er en del av Glow-strukturen. Dette laget bryter ned dataene i to deler, x_a og x_b, og anvender affine transformasjoner på dem. Affine koblingslaget bidrar til å modellere komplekse avhengigheter i dataene og forbedrer kvaliteten på den genererte lyden.

Iterativ Prosess

Prosessen med inverterbare 1x1 konvolusjoner og affine koblingslag gjentas 12 ganger. Denne iterative prosessen tillater modellen å bygge en dyptgående forståelse av dataenes struktur og øker nøyaktigheten i den endelige lydsyntesen.

Upsampled Mel-Spectrogram

Det oppsamlede mel-spektrogrammet integreres også i denne prosessen. Dataene fra mel-spektrogrammet brukes som betingelsesinformasjon og kombineres med de transformerende vektorene x_a og x_b. Dette er en kritisk del av prosessen, da det gir modellen konteksten den trenger for å generere sammenhengende tale.

Affine Transformasjoner og WN

Ved hjelp av affine transformasjoner og WaveNet-moduler (WN), finjusteres de transformerende vektorene ytterligere. Dette gjør det mulig for modellen å generere nøyaktige og naturlige lydbølger basert på mel-spektrogrammet.

Endelig Lydproduksjon

Til slutt, når transformasjonene er fullført, produserer modellen den endelige lydsekvensen z. Denne lydsekvensen er en naturlig tale som er konstruert fra de opprinnelige mel-spektrogrammene.

Gjennom denne prosessen demonstrerer WaveGlow hvordan kombinasjonen av Glow og WaveNet kan brukes til å generere høy kvalitet på tale fra mel-spektrogrammer. Den iterative prosessen med affine koblingslag og inverterbare konvolusjoner, kombinert med betingelsesinformasjonen fra mel-spektrogrammet, sikrer at den genererte lyden er både naturlig og av høy kvalitet. Dette avsnittet er inspirert av Prenger mfl., 2018.

2.3.8 Nvidia CUDA

CUDA plattformen er en softwareplattform utviklet av NVIDIA for å utvide mulighetene for GPU akselerasjon. Den tillater utviklere å bruke alle tilgjengelige ressurser på CUDA GPU-er til å prosessere data raskere enn tradisjonelle CPU-er (Priya, 2022).

2.3.9 Nvidia cuDNN

Nvidia CUDA Deep Neural Network (cuDNN) er er GPU-akselerert bibliotek for dype nevrale nettverk (Nvidia, 2024a).

2.3.10 Nvidia RIVA

Nvidia RIVA er en sett med GPU-akselererte tale- og oversettelses-tjenester for sanntid AI-pipelinjer. RIVA er designet for å settes opp i skyer, datasentere eller servere, og gir lett tilgang for brukere på alle nettverkskoblede enheter (Nvidia, 2024c).

3. Metode

Denne delen har som mål å gi en klar beskrivelse av den anvendte forskningsmetodikken, tilnærmingen og prosessen brukt for å utføre vårt arbeid på en vitenskapelig måte.

3.1 Forskningsmetode

Forskningsmetoden som ble anvendt i prosjektet, er basert på design science research (DSR)-modellen, tilpasset for å passe til dette spesifikke tilfellet. Metoden følger disse fasene:

1. Tidligere erfaring og motivasjon
2. Litteratur analyse
3. Utvikling av designkonsepter
4. Valg av teknologier
5. Valg av datasett
6. Dataforberedelse
7. Modellbygging og trening
8. Evaluering og finjustering
9. Distribusjon og anvendelse

Denne metoden er svært egnet for implementering av en TTS-modell basert på en egen stemme.

3.2 Fremgangsmetode

3.2.1 Tidligere erfaring og motivasjon

Målet med oppgaven var å benytte Tacotron 2 og WaveGlow for å konvertere tekst til tale og så implementere de trente modellene inn i systemet til Pexip ved hjelp av Nvidia RIVA server. Basert på dette, ble det opprettet mindre problemstillinger underveis for å best kunne oppnå målet. Vi delte problemstillingen opp i disse mindre problemstillingene:

- Utvikle eller benytte et allerede eksisterende norsk datasett.
- Justere modellen fra Nvidia til å passe et norsk datasett.
- Tren og overvåk treningsprosessen av Tacotron 2 og WaveGlow modellene.

3.2.2 Litteraturanalyse

Det ble utført en litteraturanalyse for å utforske tidligere arbeid som har blitt gjort på dette området. Denne litteraturgjennomgangen ble utført gjennom et omfattende søk og grundig analyse av ulike relevante akademiske arbeider, forskningsstudier og publikasjoner relatert til emnet. Dette la grunnlaget nødvendig for å løse problemstillingen.

De spesifikke spørsmålene som ble utforsket med litteraturanalysen var:

- Eksisterer det allerede liknende arbeid for andre språk?
- Hvilke utfordringer er sannsynlige å støte på med en slik oppgave?

Basert på litteraturanalysen hadde vi følgene funn:

- Det er få tilgjengelige eksempler på lignende oppgaver på nettet, hvorav de fleste er på engelsk og andre er på diverse germanske språk.
- Nesten alle disse studiene anvendte tidligere trente modeller av Tacotron 2 og WaveGlow, som var opplært på omfattende datasett før de ble tilpasset til de spesifikke behovene i hvert tilfelle.
- Det er utfordrende å trene modeller som Tacotron 2 og WaveGlow; derfor anbefales det å benytte GPU-klustere som Idun.

3.2.3 Utvikling av designkonsepter

Basert på resultatene fra litteraturanalysen, ble det besluttet å anvende Nvidias Tacotron 2 og WaveGlow-modeller fra deres GitHub-repositorium som et utgangspunkt. Dette repositoriet er designet for tekst-til-tale-konvertering ved hjelp av Tacotron 2 og WaveGlow for engelsk. Koden måtte imidlertid modifiseres for å tilpasses norsk språk og et annet datasett.

Det ble videre besluttet at det var fordelaktig å initiere treningen fra en forhåndstrent modell dersom datasettet var begrenset. Dette fremmet raskere konvergens ved bruk av mindre datasett og minsket risikoen for overtilpasning, selv om de forhåndstrente modellene opprinnelig var trent på et annet språk.

Til slutt ble det avgjort å benytte Idun GPU-kluster for treningen av modellene, gitt de utfordringer som er forbundet med å trene slike modeller uten tilsvarende ressurser.

3.2.4 Valg av teknologi og utviklingsmetode

Basert på problemstillingen fra Pexip og litteraturanalysen ble det bestemt å utnytte følgene teknologier:

- Tacotron 2: Påkrevd av Pexip.
- WaveGlow: Påkrevd av Pexip.
- Tensorboard: Verktøy for å få en god oversikt over treningsprosessen til modellene.
- Idun GPU-kluster: GPU-kluster som gjør det mulig å bruke kraftige gpuer for å trene modellene.
- Nvidia RIVA: Programvareplattform for å implementere løsningen i Pexip Infinity slik at TTS skjer i sanntid.

3.2.5 Valg av datasett

Hovedmålet med denne oppgaven var å utvikle en avansert TTS-modell som høres realistisk og menneskelig ut. For å oppnå dette var det avgjørende å ha et datasett med høy lydkvalitet, uten mye bakgrunnsstøy i lydklippene, som var av én persons stemme og med nødvendige rettigheter for kommersiell bruk. Etter omfattende søk på nettet ble det ikke funnet datasett som oppfylte alle disse kravene, så den beste løsningen ble å lage et eget datasett ved å spille inn egen stemme.

Skolen tilbyr egne podcast-rom som er utstyrt med gode mikrofoner og støydempende vegger. Grunnet tidsbegrensninger ble disse rommene ble benyttet for å ta opp deler av lydklippene til datasettet. Resten ble tatt opp med egen mikrofon.

I tillegg til høy kvalitet på lydklippene, er det viktig at modellen eksponeres for et variert utvalg av ord og sjangre av tale. For datasettet ble det brukt kilder som er tillatt for kommersiell bruk, som Wikipedia, Stortingets møter, gamle eventyr og hverdagslige samtaler. Disse kildene gir en solid grunnmur for det norske språket som modellen kan lære av.

3.2.6 Dataforberedelse

Før opplæringen av modellen påbegynnes, er det avgjørende at datasettet er formatert slik at det er velegnet for både Tacotron 2 og WaveGlow. Det første trinnet var å konvertere lydopptakene til et passende format, i tråd med det som anvendes av den eksisterende engelske modellen som bruker LJSpeech-datasettet. Dette innebærer lydklipp med en samplingsfrekvens på 22050 Hz i mono.

Datasettet ble deretter renset for uønskede tegn og symboler som kunne forstyrre treningsprosessen.

Videre ble datasettet delt i tre deler: et treningssett, et valideringssett og et testsett. Treningssettet inneholder data som brukes til opplæring av modellen, valideringssettet benyttes for å justere modellens arkitektur og finjustere hyperparametere, mens testsettet anvendes for en endelig evaluering av modellens ytelse etter fullført trening og validering. Datasettet ble delt inn i 85% for trening, 10% for validering og 5% for testing, en fordeling som er vanlig og bidrar til å utvikle en robust og godt generaliserbar modell.

Til slutt måtte det endelige datasettet formatteres spesifikt for at modellen skulle kunne identifisere hvilken tekst som tilhører hvilket lydklipp eller mel-spektrogram (2.1.1). Dette ble oppnådd ved å bruke CSV-filer hvor første kolonne angir filstien til hvert lydklipp, og den andre kolonnen inneholder teksten som leses opp i det tilhørende lydklippet. Dette formatet ble anvendt både for lydfilene og mel-spektrogrammene.

Alt i alt, bestod datasettet av 10 timer total tale hvorav fem timer var originalopptak og fem timer var databeriket data. Av disse fem timene var 1 time tatt opp på podcast-rom.

3.2.7 Modellbygging og trening

Tacotron 2 og waveglow er begge modeller som er i stor grad språkuavhengige. Videre er engelsk og norsk begge germanske språk med mange likheter. Det betyr at man ikke trenger å foreta mange endringer i modellarkitekturen for å tilpasse den et annet språk.

Videre, ble bestemt å bruke en forhåndstrent modell for både Tacotron 2 og WaveGlow som utgangspunkt til treningen. Disse modellene er utviklet av Nvidia og er trent på et stort engelsk datasett brukt for å trene en engelsk tts-modell. Ved å bruke de trente vektene til disse modellene, konvergerer modellene betydelig raskere samtidig som det forebygger overtilpasning.

For å sikre konsistens i treningsprosessen ble tidligere trente modeller benyttet som grunnlag. Det er derfor essensielt at arkitekturen til modellene som er trent på det engelske datasettet, og arkitekturen som brukes for det norske datasettet, er identiske. En viktig komponent i modellarkitekturen er antallet symboler som benyttes i datasettet. For engelsk er dette tallet 148, inkludert alle symboler, bokstaver og -representasjoner, mens for norsk ville dette tallet være 154 med de ekstra bokstavene i alfabetet. For å harmonisere arkitekturene, ble en metode utviklet for å konvertere de norske bokstavene æ, ø, å, Æ, Ø og Å til deres engelske representa-

sjoner, for eksempel blir Å til Aa. Dette blir gjort automatisk og påvirker ikke formatet til datasettet eller bruk av modellene i praksis.

3.2.8 Ytelsesvurdering og optimalisering

Etter grunnleggende implementeringer av prioriterte tilnærmingene, ble fasen for ytelsesevaluering og optimalisering satt i gang. I denne ble verktøyet Tensorboard utnyttet for å overvåke treningsprosessen kontinuerlig. Tensorboard registrerte treningstap, valideringstap og læringsraten etter hver epoke, hvilket gjorde det mulig å visualisere modellens fremgang under treningen og lettere velge en passende læringsrate.

I tillegg ble forskjellige varianter av tilnærmingene testet gjennom ytelsesevaluering. Basert på innsamlede data fra disse testene ble det gjort justeringer for å undersøke mulighetene for optimalisering av tilnærmingene. Dette omfattet testing av diverse hyperparametere og kombinasjoner av testdata. Vi anvendte også ulike kombinasjoner av teknikker innenfor hver tilnærming, og justerte modell-arkitekturparametre basert på datasettets størrelse for å finjustere treningsprosessen.

4. Resultater

4.1 Vitenskapelige resultater

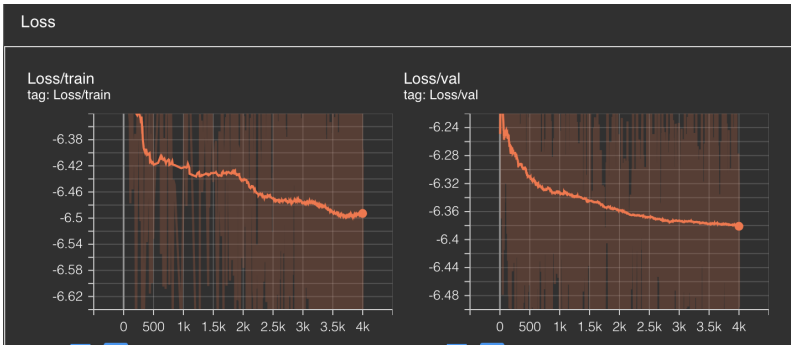
Prosjektet hadde som mål å utvikle en tekst-til-tale modell som produserer syntetisk norsk tale basert på norsk tekst. For å oppnå dette, utviklet vi et norsk datasett på 10 timer som ble brukt til å finjustere ferdigtrente modeller av Tacotron 2 og WaveGlow. Ved å bruke disse forhåndstrente modellene som utgangspunkt, kunne vi effektivt bygge videre på allerede trente modell-vekter, noe som førte til betydelig bedre resultater sammenlignet med å trene fra bunnen av.

Treningsparametere og treningsprosess: Følgende treningsparametere ble brukt for modellene:

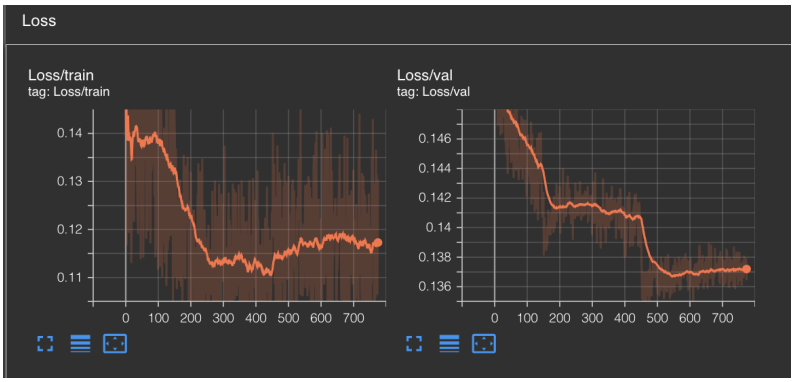
Modell	Læringsrate	Batch størrelse	Antall epoker	Vekttap
Tacotron 2	0.0005	40	1000	0.000001
WaveGlow	0.00001	10	4000	0

Treningen av Tacotron 2 og WaveGlow ble utført hver for seg. Tacotron 2 bruker betydelig lenger tid å trene enn WaveGlow, så det er viktig å overvåke treningsprosessen underveis for å forebygge overtilpassing til datasettet.

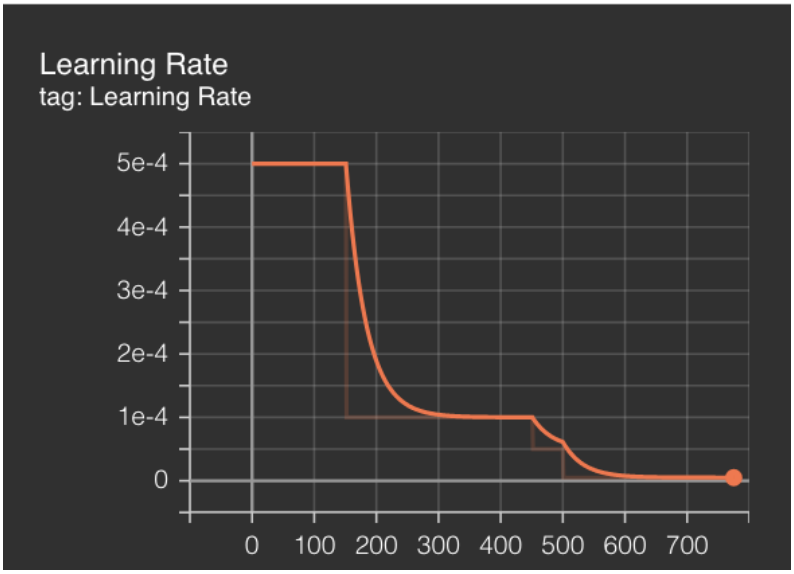
Loss-grafer: Under treningen av modellene målte vi tap (loss) for å overvåke ytelsen og forbedringene. Figur 4.1 til 4.4 viser trenings- og valideringskurvene for Tacotron 2 og WaveGlow modellene samt treningsraten som ble brukt gjennom trening. Disse grafene viser en klar nedgang i tap over tid, noe som indikerer at modellene lærte å produsere mer nøyaktig syntetisk tale.



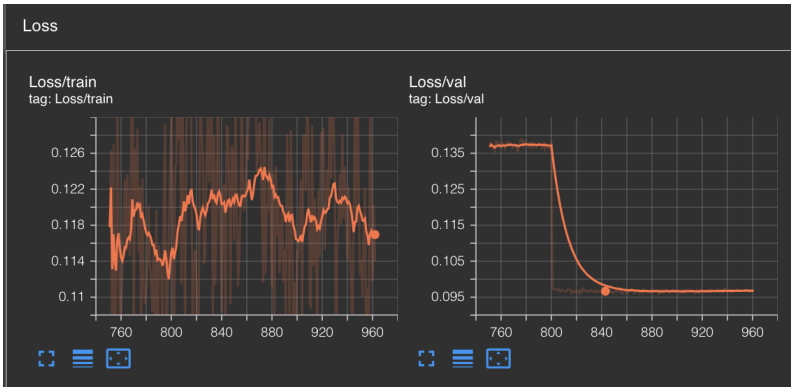
Figur 4.1: WaveGlow sin tapsgraf etter 4000 epoker. Læringsraten ble holdt stabil på 0.0004 gjennom trening.



Figur 4.2: Tacotron 2 sin tapsgraf etter 800 epoker.



Figur 4.3: Utviklingen av treningsraten til Tacotron 2 gjennom trening.

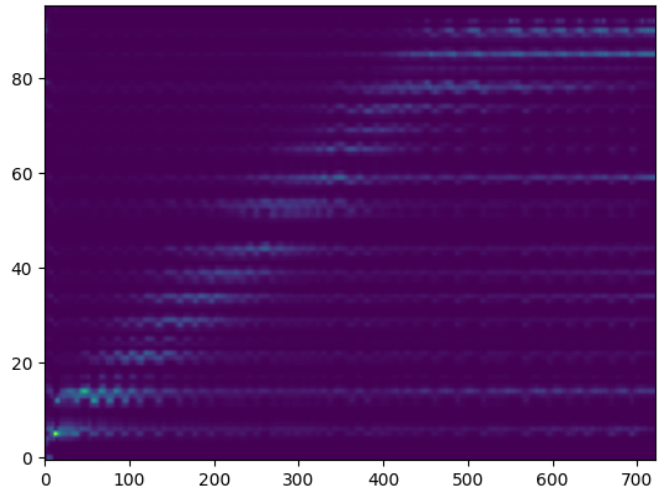


Figur 4.4: Tacotron sin tapsgraf fra 700 til 1000 epoker. Den bratte nedgangen av valideringstap skyldes at den begynte å trene uten databeriket lydklipp.

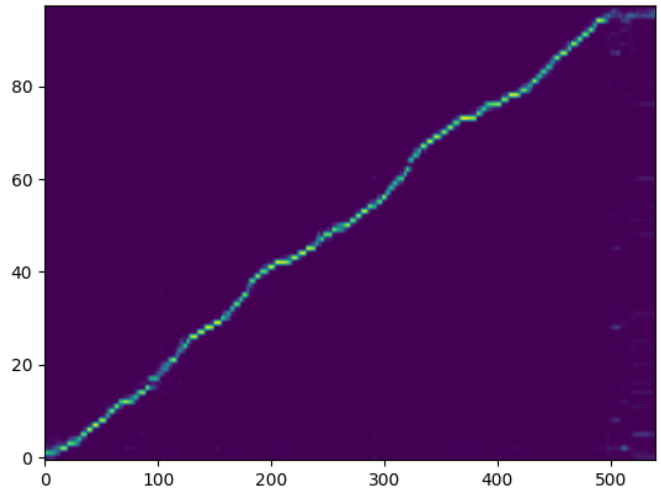
Som grafen i figur 4.3 viser ble læringsrate endret underveis. Den begynte på en høyere verdi og ble gradvis lavere. Læringsraten ble endret når prosessen viste tegn til overtilpassing eller lite fremgang. Dette hjelper modellen konvergere raskere.

Mot slutten av treningen begynte modellen å generere ustabile resultater, selv om den syntetiserte samme inntekst. Dette ble løst ved å sette en fast seed for treningen av Tacotron 2 og WaveGlow. I tillegg ble datasettet som Tacotron 2 trenes på endret til å kun inkludere de originale lydklippene, ettersom de justerte lydfilene kunne ha forstyrret treningen ved at modellen ikke visste hvordan den skulle generere mel-spektrogrammene. Denne endringen er tydelig i figur 4.4, hvor valideringstapet sank betydelig etter fjerning av de databerikede lydfilene.

Alignment-grafer: Modellens evne til å matche tekst med tale ble evaluert ved bruk av alignment-grafer. Figur 4.5 og 4.6 viser eksempler på alignment-grafer før og etter bruk av ferdigtrente modeller på den samme innteksten. Modellene som ble trent på forhåndstrente modeller, viser et tydeligere og mer sammenhengende diagonalt mønster, noe som indikerer bedre justering mellom inputtekst og generert tale.



Figur 4.5: Alignment graf uten bruk av forhåndstrente modeller som utgangspunkt.



Figur 4.6: Alignment graf etter bruk av forhåndstrente modeller som utgangspunkt.

4.1.1 Evaluering av den syntetiske talen

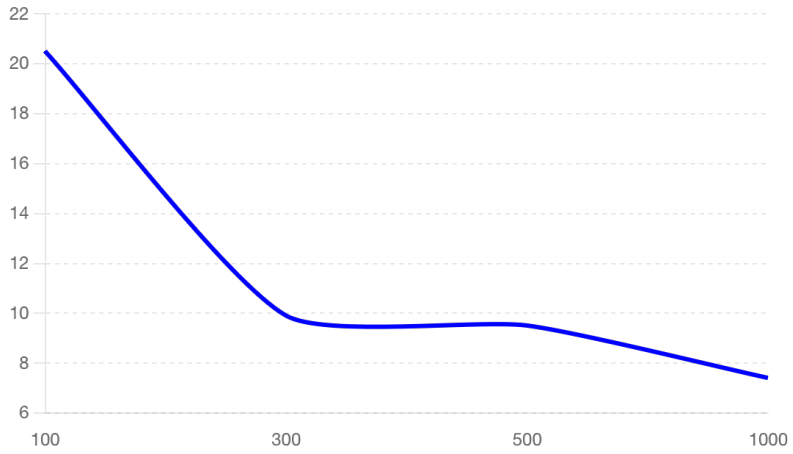
For å avgjøre om målet om å lage en tekst til tale modell som høres naturlig ut ble oppfylt, ble det utført forskjellige tester for å evaluere kvaliteten

av den syntetiske talen. Målingene nedenfor ble beregnet ved å ta gjennomsnittet av fem ulike målinger av fem forskjellige setninger. Dette gir mer robuste resultater.

Mel-Cepstral Distortion (MCD): måler forskjellen mellom de originale og de syntetiske mel-cepstral-koeffisientene. Det er en vanlig brukt metrikk for å evaluere kvaliteten på syntetisk tale. En skår på 1-2 er utmerket og veldig vanskelig å skille fra den originale talen, 5-6 er fortsatt bra, men det er en merkbar forskjell mellom originalen og den syntetiserte talen. Alt over 8 er stor forskjell fra originalen. Resultatene av MCD-målingene er vist i figur 4.1.

Epoke Tacotron 2	Epoke WaveGlow	MCD verdi
100	4000	20.5
300	4000	9.9
500	4000	9.5
1000	4000	7.4

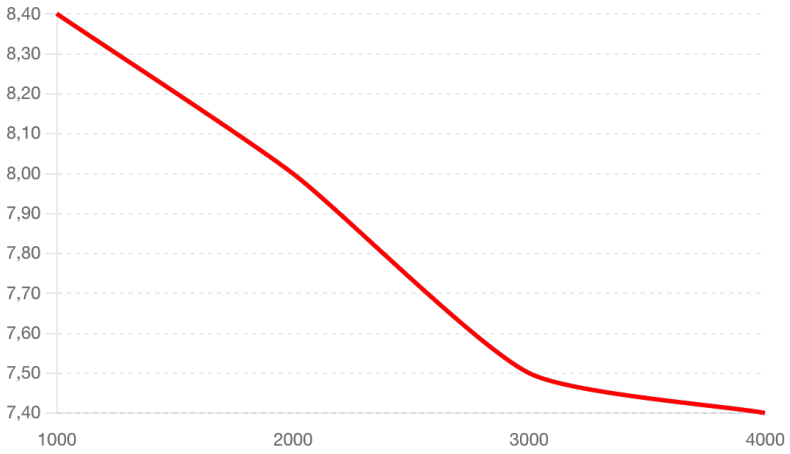
Tabell 4.1: Oversikt over hvordan trening av Tacotron 2 påvirker modellens MCD-verdier.



Figur 4.7: Visuell representasjon av tabellen i 4.1 der x-aksen er antall epoker og y-aksen er MCD-verdi.

Epoke Tacotron 2	Epoke WaveGlow	MCD verdi
1000	1000	8.4
1000	2000	8.0
1000	3000	7.5
1000	4000	7.4

Tabell 4.2: Oversikt over hvordan trening av WaveGlow påvirker modellens MCD-verdier.



Figur 4.8: Visuell representasjon av tabellen i 4.2 der x-aksen er antall epoker og y-aksen er MCD-verdi.

Disse resultatene viser gjennomsnittet av fem målinger på fem forskjellige innputter. Under beregningene ble det tydelig at modellens ytelse varierte mellom ulike tekster. Derfor ble det gjennomført en grundigere undersøkelse av hvor godt den beste modellen presterte på flere lydklipp. Modellen ble evaluert på 20 forskjellige lydklipp, med en gjennomsnittlig skår på 6.8, en laveste verdi på 5.3, og en høyeste verdi på 9.0.

Meningsmålinger (MOS): MCD er en god indikator for å evaluere kvaliteten på syntetisert tale, men den er ikke perfekt. Talen kan variere i tonefall, intonasjon eller prosodi og derfor få en høy MCD-verdi, men fremdeles være realistisk. Det er derfor viktig å supplere MCD med andre metoder som MOS. For å teste hvor godt den syntetiske talen høres ut, ble det utført en spørreundersøkelse der deltakerne kunne vurdere kvaliteten av talen. Prosessen innebærer at en gruppe testpersoner lytter til et sett med lydprøver og gir dem en vurdering på en skala fra 1 til 5, hvor:

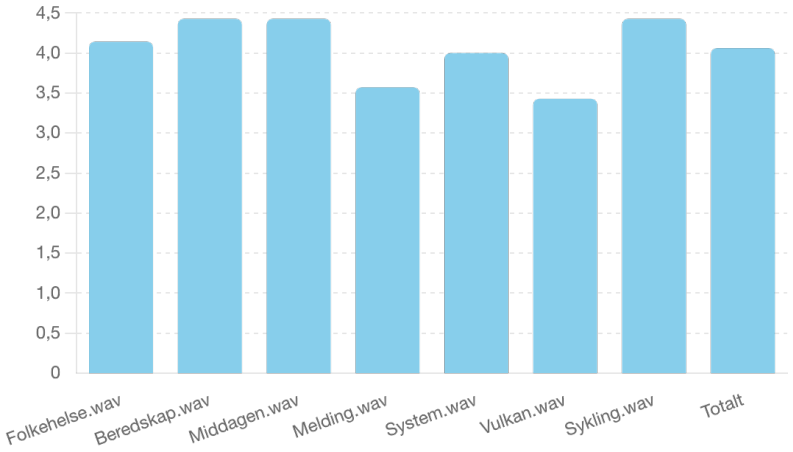
- 5: Utmerket – ingen merkbar forringelse i lydkvaliteten.
- 4: God – liten grad av forringelse som er merkbar, men ikke forstyrrende.
- 3: Middels – merkbar forringelse som kan være litt forstyrrende.
- 2: Dårlig – forringelse er merkbar og forstyrrende.
- 1: Svært dårlig – forringelse er veldig merkbar og svært forstyrrende.

Testen ble utført av sju deltakere, som ble bedt om å vurdere sju forskjellige lydklipp. Disse lydklippene var generert av den beste versjonen av systemet og inkluderte både data som systemet hadde blitt trent på, og data som det ikke hadde blitt trent på. Resultatene av testen er som

følger:

Lydklipp	deltaker 1	deltaker 2	deltaker 3	deltaker 4	deltaker 5	deltaker 6	deltaker 7
Folkehelse.wav	4	4	5	4	4	4	4
Beredskap.wav	5	4	4	5	5	4	4
Middagen.wav	5	5	3	5	4	5	4
Melding.wav	4	3	4	4	3	3	4
System.wav	4	3	5	4	4	4	4
Vulkan.wav	3	2	4	4	4	4	3
Sykling.wav	5	5	5	4	4	4	4
Totalt	4.29	3.71	4.29	4.29	4.00	4.00	3.86

Tabell 4.3: Gjennomsnittlig kvalitetsskåring for hver deltaker



Figur 4.9: Visuell representasjon av meningsmålinger der hver søyle representerer et lydklipp og y-aksen representerer den gjennomsnittlige skåren lydklippet fikk.

Resultatene viser tydelig forbedring i kvaliteten på den syntetiske talen før og etter bruk av forhåndstrente modeller. Før vi implementerte forhåndstrente modeller, var den syntetiske talen unaturlig og hadde flere feil i uttale og intonasjon.

Taps-grafene indikerer også at WaveGlow konvergerte raskere enn Tacotron 2. Dette er logisk, siden WaveGlow trenes på én time med høy kvalitet tale, sammenlignet med de ti timene Tacotron 2 krever.

Evalueringen av den syntetiske talen viser at modellen forbedret seg betydelig gjennom treningsprosessen, med de beste resultatene fra de mest veltrente modellene.

Meningsmålingen viser også tilfredsstillende resultater med en gjennomsnittlig skår på omtrent 4 av 5. Dette indikerer at talekvaliteten er god for de fleste deltakere.

4.2 Ingeniørfaglige resultater

I starten av prosjektet ble det satt flere mål, inkludert utvikling av et norsk datasett, integrasjon og finjustering av Tacotron 2 og WaveGlow, samt evaluering av den endelige modellen gjennom ulike tester. Status for disse målene er som følger:

- **Utvikling av norsk datasett:** Vi samlet og bearbeidet et 10-timers datasett bestående av norsk tale og tilsvarende tekst. Datasettet ble nøye kontrollert for kvalitet, inkludert sjekk for bakgrunnsstøy og uttalefeil.
- **Integrasjon og finjustering av modeller:** Vi brukte de forhåndstrente modellene Tacotron 2 og WaveGlow som utgangspunkt og finjusterte dem med vårt datasett. Finjusteringen forbedret modellens ytelse betydelig.
- **Evaluering:** Den endelige modellen ble evaluert gjennom både subjektive og objektive metoder. Subjektive tester inkluderte lytteprøver hvor deltakerne vurderte naturligheten av den syntetiske talen. Objektive tester inkluderte beregninger av metrikker som mel-spectrogram avstand.

Ved leveringstidspunktet fungerer systemet tilfredsstillende og kan produsere norsk tale fra tekstinput. Systemet har gjennomgått flere testrunder som viser forbedringer i både klarhet og naturlighet av talen sammenlignet med baseline-modellene, selv om det tidvis kan gi ustabile resultater. Resultatene viser jevne forbedringer gjennom treningsfasen og en akseptabel ytelse i testene.

4.3 Administrative resultater

Prosjektet fulgte en strukturert fremdriftsplan med GANTT-diagrammet (vedlagt). Her er en oppsummering av måloppfyllelsen i forhold til framdriftsplanen:

- **Fremdriftsplan:** Prosjektet fulgte den opprinnelige planen tett, med noen justeringer basert på innsamling av data og modelltrening. De fleste milepælene ble nådd innenfor de estimerte tidsrammene.
- **Timeregnskap:** Totalt ble det brukt 600 timer på prosjektet, fordelt på aktiviteter som datainnsamling, modelltrening, evaluering, og dokumentasjon. Dette inkluderte også tid brukt på møter og koordinering.
- **Utviklingsprosess:** Det ble holdt ukentlige møter med arbeidsgiver

gjennom mesteparten av prosessen. Gruppemedlemmene holdt kontakt stadig gjennom prosessen med obligatoriske møter hver uke.

Samlet sett har prosjektet oppnådd sine mål ved å utvikle en høyytelses TTS-modell for norsk tale, og resultatene viser en klar forbedring i kvaliteten på syntetisk tale etter bruk av forhåndstrente modeller.

5. Diskusjon

I denne diskusjonsdelen vil resultatene av utviklingen av en norsk tekst-til-tale-modell ved hjelp av Tacotron 2 og WaveGlow bli analysert. Hovedresultatene fra kapittel 4 vil bli kort oppsummert, før en detaljert drøfting av disse følger.

5.1 Analyse av Resultater

5.1.1 Sammenligning med Mål og Planer

De opprinnelige målene var å utvikle en modell som kunne syntetisere norsk tale med høy kvalitet. Resultatene viste at mens noen aspekter oppfylte disse målene, var det også avvik. For eksempel var modellens kvalitet inkonsistent avhengig av inndata. Dette avviket skyldes sannsynligvis et for lite norsk datasett med begrenset ordforråd.

5.1.2 Oppfyllelse av Krav

Sluttproduktet oppfylte kravet om å utvikle en tekst til tale modell som høres naturlig ut. Med en gjennomsnittlig skår på 4/5 i meningsmålingen kan man konkludere at modellen gjør jobben sin.

5.1.3 Sluttprodukt og Oppdragsgivers Forventninger

Sluttproduktet er en fungerende tekst-til-tale-modell som ofte gir resultater av høy kvalitet, men som kan til tider generere unaturlig stemme. Basert på MCD og MOS resultatene kan man konkludere med at talen som genereres ofte gir en liten forringelse som er merkbar, men ikke forstyrrende.

5.2 Evaluering av Prosess, Fremgangsmåte og Teknologi

5.2.1 Positivt Resultat Grunnet Prosess og Teknologi

En stor fordel ved å bruke Tacotron 2 og WaveGlow for å syntetisere tekst til tale er den omfattende tilgjengelige informasjonen på området. Begge modellene er relativt etablerte, noe som betyr at mange allerede har

utviklet tekst-til-tale-systemer for engelsk og noen for andre ulike språk. Dette ga et solid grunnlag for å løse oppgaven. For eksempel ble ideen om å bruke en forhåndstrent modell som utgangspunkt, før modellene trenet på det norske datasettet, inspirert av andre på nettet. Dette forbedret modellens ytelse betydelig.

5.2.2 Negativt Resultat Grunnet Prosess og Teknologi

Til tross for bruken av forhåndstrente modeller som utgangspunkt for treningen av modellene, er et datasett på 10 timer fortsatt begrenset. Videre bestod kun halvparten av datasettet av unike setninger, ettersom databerikelse ble benyttet for å øke størrelsen. Dette kan ha resultert i et begrenset ordforråd og kan forklare hvorfor resultatene kan være inkonsistente til tider. I tillegg var kun én time av datasettet av tilstrekkelig kvalitet for å trene WaveGlow-modellen. Dette kan føre til inkonsistent tale i visse tilfeller og er et tydelig forbedringsområde.

Anvendelsen av Tacotron 2 og WaveGlow som grunnlag for tekst-til-talesyntese kan ha ytterligere påvirket resultatene negativt. Begge disse modellene, utviklet henholdsvis i 2017 og 2018, kan anses som utdaterte i dagens hurtig utviklende teknologilandskap. Tacotron 2, introdusert i desember 2017, kombinerer en sekvens-til-sekvens modell for generering av mel-spektrogrammer med en WaveNet-lignende modell for bølgeformsyntese. WaveGlow, publisert i desember 2018, benytter en kombinasjon av Glow-modellen for sannsynlighetstetthet estimering og WaveNet for bølgeformgenerering.

Til tross for at begge modellene representerte fremskritt på tidspunktet for sin lansering, har nyere metoder overgått dem både i ytelse og effektivitet. Nyere modeller, som for eksempel FastSpeech og VITS (Variational Inference Text-to-Speech), tilbyr betydelige fordeler. Disse nyere teknologiene krever ofte mindre omfattende datasett, men leverer likevel høyere lydkvalitet og raskere prosesseringstider takket være forbedrede arkitekturer og avanserte læringsteknikker.

På grunn av disse teknologienes alder og de observerte begrensningene i datasettet, er det sannsynlig at bruken av Tacotron 2 og WaveGlow ikke var de optimale modellene for denne oppgaven. Nyere modeller har potensial til å levere mer naturlig og konsistent tale med mindre datasett, noe som kunne ha resultert i bedre ytelse og mer tilfredsstillende resultater.

5.3 Styrker og Svakheter ved Resultatene

5.3.1 Styrker

Resultatene viser lite forskjell mellom taps-grafene for trening og validering. Dette tyder på at man kan forvente like gode resultater på data modellen har trent på og data modellen ikke har sett før.

5.3.2 Svakheter

Resultatene for MCD ble mål på trent data fordi denne metoden krever et original lydklipp å sammenlikne med det syntetiserte lydklippet. Det betyr at modellen ikke blir objektivt mål på usett data. Selv om taps-grafene for validering og trening er relativt like og spørreundersøkelsen inneholdt utrent data, kunne det likevel vært hensiktsmessig å målt dette med en objektiv metode som for eksempel Word Error Rate (WER) som sammenlikner innteksten med en transkribert versjon.

5.4 Anbefalinger for Videre Arbeid

Videre forskning bør fokusere på å øke størrelsen på datasettet med høykvalitets lydklipp. Dette vil både hjelpe Tacotron 2 med å håndtere utfordringer knyttet til ordforråd og forbedre WaveGlows uttale av setninger.

Det er også anbefalt å undersøke mer moderne modeller som kan prestere bedre på et mindre datasett.

5.5 Refleksjon over Egen Arbeidsinnsats og Læring

Dette prosjektet har gitt verdifull innsikt i dyp læring og tekst-til-tale-teknologier. Utfordringer som oppstod underveis, spesielt knyttet til datainnsamling og modelltrening, har forbedret ferdighetene i problemløsning betydelig.

I starten ble modellene trent på en egen PC, noe som viste seg å være uhensiktsmessig gitt prosjektets kompleksitet og ressursbehov. De fleste problemene i prosjektet var relatert til et lite datasett, og det tok betydelig tid å øke datasettets størrelse gjennom databerikelse og mange timer med innspilling av ekstra lydklipp. Mange av de endringene og justeringene som ble gjort før datasettstørrelsen ble økt, og før bruken av forhåndstrente modeller som utgangspunkt, kan anses som bortkastet tid.

Til slutt ble det utviklet en fungerende modell med tilfredsstillende ytelse. Likevel, med dagens kunnskap, ville en ny tilnærming med større fokus på innspilling av lydklipp fra starten av ført til en vesentlig bedre modell. Erfaringene fra dette prosjektet har derfor vært uvurderlige for forståelsen av effektiv modelltrening og viktigheten av et omfattende og variert datasett.

6. Konklusjon og videre arbeid

Denne oppgaven har presentert hvordan man kan lage Text-til-tale modeller på brukerens ønskede språk. Samtidig som om at det også blir presentert en modell på Norsk med god kvalitet. Datasettet brukt er bygget opp fra grunnen av, og om man ønsker å videreføre det er det fult mulig. Siden det ikke finnes gode store datasett med enkeltpersoner som snakker er det vanskelig å produsere en god Text-til-tale modell på språket det gjelder.

For videre utvikling av modelllen anbefales det å jobbe enda mer med samling av datasett for å kunne øke kvaliteten, og kunne øke kompleksiteten på modellen slik at den kan håndtere så mange vanskelige ord å setninger som mulig. Dete gjelder spesielt for Tacotron 2, da denne krever mye data og lang treningstid.

I den originale oppgaven var det et ønske om å få modellen til å kunne kjøre lokalt. Dette ble desverre ikke sett på da prioritering av tid og resurser måtte fokuserer på problemene som oppsto og ble diskutert i 5.2.2. Lokal kjøring burde være mulig å gjennomføre ved et senere arbeid om det er ønskelig, enten som en ny oppgave, eller av Pexip selv.

Forhåpentligvis legger denne rapporten frem gode punkter for implementasjon av egne Text-til-tale modeller. Samtidig som at den leverer et tilfredstillende produkt som kan brukes og videreutvikles av Pexip etter endt arbeid.

Samfunnspåvirkning

Målet med oppgaven var å lage en TTS-modell som kan anvendes i Pexip sine applikasjoner for lett tilgjengelighet og bruk av deres brukere. Det kan gjøre det lettere for personer med nedsatt lese eller talevansker å få med seg ting som skjer i møter og samtaler, eller åpner opp muligheter å få stemmen sin hørt i et møte, om man er bekymret for eller ikke klarer å ta ordet. TTS finnes i mange former allerede men flere av dem er låst bak betalingsvegger, og dårlige modeller, spesielt for mindre språk.

Samfunnsmessig har denne teknologien et stort potensiale for å jevne ut store forskjeller i samfunnet. Det å ha lett tilgjengelige hjelpemidler for kommunikasjon hjelper til med å utgjevne språkforskjeller. Samtidig har utviklingen til teknologien også skapt bekymringer for at teknologien kan bli misbrukt. En populær nå måte å missbruke dette på, er at pesoner som prøver å lure deg, kan ta stemmen til familiemedlemmer, eller venner, og ringe offeret med deres stemme og spørre om penger (Cerullo, 2023). Om dette blir mer vanlig kan det påvirke hvordan man kan stole på hverandre. Ens stemme har i lang tid vært en unik og gjenkjennbar identitet, og om det blir gjett å jukse frem må man begynne å tvile mer på hverandre.

Det å trene store maskinlæringsmodeller krever mye energi. Utviklingen i datasentere krever mer og mer strøm, og det er viktig at man tenker på dette før man setter igang store prosjekter. Heldigvis er det mulig med teknologien anvendt i oppgaven mulig å forbedre og utvikle fra tidligere modeller, selv på andre språk. Dette reduserer kravene for trening betraktelig, og betyr at utviklingen krever betydelig mindre energi enn om man skal begynne fra bunnen av.

Bibliografi

Banoula, M. (2023 april). Introduction to Long Short-Term Memory(LSTM). https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/lstm

Bayram Durna, M. (2024 januar). Exploring Seq2Seq, Encoder-Decoder, and Attention Mechanisms in NLP: Theory and Practice. https://medium.com/@mervebdurna/exploring-seq2seq-encoder-decoder-and-attention-mechanisms-in-nlp-theory-and-practice-9b1022cf50b4

Cerullo, M. (2023 mars). AI scams mimicking voices are on the rise. https://www.cbsnews.com/news/ai-scam-voice-cloning-rising/

Docker. (2024 mars). What is a Container? | Docker. https://www.docker.com/resources/what-container/

Doshi, K. (2021 februar). Audio Deep Learning Made Simple - Why Mel Spectrograms perform better. https://ketanhdoshi.github.io/Audio-Mel/#:~:text=A%20Mel%20Spectrogram%20makes%20two,of%20Amplitude%20to%20indicate%20colors.

Gartzman, D. (2019 august). Getting to Know the Mel Spectrogram - Towards Data Science. https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

IBM. (2024 april). What Is Machine Learning (ML)? | IBM. https://www.ibm.com/topics/machine-learning

Intel. (2020). Convolutional Neural Networks (CNN) and Deep Learning. https://www.intel.com/content/www/us/en/internet-of-things/computer-vision/convolutional-neural-networks.html

Nvidia. (2024a). CUDA Deep Neural Network. https://developer.nvidia.com/cudnn

Nvidia. (2024b). NVIDIA Container Toolkit | NVIDIA NGC. https://catalog.ngc.nvidia.com/orgs/nvidia/teams/k8s/containers/container-toolkit

Nvidia. (2024c). NVIDIA Riva. https://www.nvidia.com/en-us/ai-data-science/products/riva/

Prenger, R., Valle, R., & Catanzaro, B. (2018 oktober). *WAVEGLOW: A FLOW-BASED GENERATIVE NETWORK FOR SPEECH SYNTHESIS*. https://arxiv.org/pdf/1811.00002

Priya, B. (2022 november). Understanding NVIDIA CUDA: The Basics of GPU Parallel Computing. https://www.turing.com/kb/understanding-nvidia-cuda

Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R., Agiomyrgiannakis, Y., & Wu, Y. (2017 desember). *NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS*. https://arxiv.org/pdf/1712.05884v1

Tidemann, A., & Elster, A. C. (2023 juli). maskinlæring – Store norske leksikon. https://snl.no/maskinl%C3%A6ring

A. Forprosjektplan

Non-English Text-to-Speech synthesis for video conferencing assistant
Forprosjektplan

Fredrik Bache Ruud
Martin Almenning

Innholdsfortegnelse

1. Mål og rammer

1.1. Orientering.

Denne oppgaven var en av flere ferdigdefinerte prosjektoppgaver som ble gjort tilgjengelige for oss. Vi ble spesielt tiltrukket av denne oppgaven på grunn av dens direkte relevans til fremtidens teknologiske landskap.

Denne oppgaven var spesielt interessant med tanke på at vi får mulighet til å utforske og anvende "state of the art" teknologi innen Text-To-Speech (TTS). Nvidia WaveGlow, som representerer et betydelig fremskritt innen dette feltet, fanget spesielt vår oppmerksomhet. Vi ser stor verdi i å forstå og anvende denne avanserte teknologien for å skape realistiske stemmer gjennom maskinlæring.

Som en gruppe har vi en felles lidenskap for maskinlæring. Hver av gruppemedlemmene har uttrykt et sterkt ønske om å fordype seg i dette spennende og raskt utviklende feltet. Ved å velge denne oppgaven, har vi muligheten til ikke bare å utvikle våre tekniske ferdigheter, men også å bidra til et felt som vi mener vil ha stor betydning i årene som kommer.

1.2. Problemstilling / prosjektbeskrivelse og resultatmål

Hovedmålet er å utvikle en TTS modell som effektivt kan omforme skrevet norsk tekst til realistisk, muntlig uttalelse. Dette skal oppnås ved å skape en stemme som høres naturlig ut, og som kan tilpasse seg ulike typer tekster.

Ved prosjektets avslutning forventes det at vi har utviklet en fullt funksjonell TTS-modell som kan integreres sømløst i Pexip sitt eksisterende system. Modellen skal være i stand til å lese og tolke norsk tekst på en måte som både er klar, naturlig, og lett forståelig for sluttbrukerne.

1.3 Effektmål

Pexip planlegger å bruke prosjektet i forbindelse med videomøter der de har en modell allerede som oversetter og lager et sammendrag av et møtetranscript. Over ett langsiktig mål ønskes det at TTS- modellen blir iverksatt her sånn att sammendraget kan bli lest opp. Dette vil betydelig forbedre tilgjengeligheten og brukervennligheten for slike språkbrukere.

Modellen skal helst også kunne kjøres lokalt av brukere med egne maskiner. Dette hjelper til med å eliminere dyre serverkostnader, som mange andre moderne maskinlæringsmodeller ofte krever.

Dette prosjektet er ikke bare en mulighet til å bidra med vår ekspertise, men også en sjanse til å videreutvikle våre ferdigheter i et meget aktuelt og fremtidsrettet område.

1.4 Rammer

Gruppen kan få behov for å låne bedre grafikkort, dette kan gjøres ved å for eksempel bruke idun for å trene modellen.

Tidsrammen til prosjektet er på omtrent fire måneder med mer detaljerte rammer beskrevet i vedlagt tidsplan.

2. Organisering

Gruppen består av to tredje-år dataingeniørstudenter som er ansvarlige for utvikling av produktet og hele prosessen rundt utviklingen av produktet. Gruppemedlemmene får veileding underveis av Tomas Holt som jobber hos NTNU, og Anna Kim og Knut Inge Hvidsten som jobber i Pexip.

3. Gjennomføring

3.1. Hovedaktiviteter

Opplisting av hovedaktiviteter:

- Research
  - Lese seg opp på Tacotron 2 og Nvidia WaveGlow for å bygge forståelse for datamodellen og TTS modeller.
  - Undersøke hva som oppfattes som en «hørbar» og god stemme til bruk i egen modell.
- Produksjon
  - Lage grunnleggende TTS modeller
  - Lage en offline TTS modell
- Dokumentasjon
  - Brukertesting
  - Rapport
  - Poster
  - Videopresentasjon

3.2. Milepæler

Opplisting av kritiske datoer.

Dato	**Hendelse**
26.01.2024 kl. 23:59	Innlevering av forprosjektplan
18.03.2024 kl. 08:15	Lage og presentere poster
21.05.2024 kl. 12:00	Prosessdokumentasjon
24.05.2024 kl. 12:00	Videopresentasjon

4. Oppfølging og kvalitetssikring

4.1 Kvalitetssikring.

Gruppen skal være i konstant kontakt med arbeidsgiver som gir veiledning dersom produktet ikke er som forventet.

Dokumentasjonen blir lagret på OneDrive for å sikre god orden. Begge medlemmene har tilgang til dokumentasjonen når nødvendig på OneDrive.

4.2 Rapportering.

Gruppen har møte med arbeidsgiver torsdag annenhver uke og med veileder der gruppen ser behov for det.

5. Risikovurdering

Dette avsnittet vil identifisere og analysere potensielle risikoer knyttet til utviklingen av TTS modellen. Målet er å vurdere sannsynligheten og potensielle konsekvenser av hver risiko, samt å foreslå tiltak for å håndtere eller minimere disse risikoene.

1. **Teknologisk usikkerhet:**
   - **Beskrivelse:** Risiko knyttet til teknologiens kompleksitet, spesielt med Nvidia WaveGlow.
   - **Sannsynlighet:** Middels
   - **Konsekvens:** Høy
   - **Tiltak:** Kontinuerlig teknologisk veiledning og regelmessig opplæring i relevante teknologier.

2. **Ressursbegrensninger:**
   - **Beskrivelse:** Potensiell mangel på nødvendige ressurser som grafikkort.
   - **Sannsynlighet:** Lav
   - **Konsekvens:** Middels
   - **Tiltak:** Tidlig forespørsel om nødvendige ressurser og backup-plan for alternativ utstyr.

3. **Gruppemedlemmenes tilgjengelighet:**
   - **Beskrivelse:** Risiko for forsinkelser grunnet medlemmenes tilgjengelighet.
   - **Sannsynlighet:** Middels
   - **Konsekvens:** Middels
   - **Tiltak:** Regelmessige møter for å synkronisere fremgang og tidsplan og sørge for at man kan jobbe hjemmefra om nødvendig.

4. **Endringer i prosjektomfanget:**
   - **Beskrivelse:** Endringer i oppgavens krav eller mål underveis.
   - **Sannsynlighet:** Lav
   - **Konsekvens:** Høy
   - **Tiltak:** Fleksibel planlegging og regelmessig kommunikasjon med veiledere og arbeidsgiver.

5. **Tekniske utfordringer med modellutvikling:**
    - **Beskrivelse:** Utfordringer knyttet til utvikling av en funksjonell og effektiv TTS-modell.
    - **Sannsynlighet:** Høy
    - **Konsekvens:** Høy
    - **Tiltak:** Forhåndstesting av modellen, regelmessig veiledning og teknisk støtte.

6. **Brukertesting og tilbakemeldinger:**
    - **Beskrivelse:** Risiko for utilstrekkelig eller negativ tilbakemelding fra brukertesting.
    - **Sannsynlighet:** Middels
    - **Konsekvens:** Middels
    - **Tiltak:** Planlegge flere runder med brukertesting og være forberedt på iterativ forbedring.

Denne risikoanalysen fokuserer på å identifisere og planlegge for potensielle hindringer i prosjektet. Gjennom å være proaktiv og ha en plan for hver identifisert risiko, kan teamet bedre navigere i prosjektet og øke sjansene for suksess.

6. Vedlegg

6.1 Tidsplan

Se vedlagt tidsplan

6.2 Adresseliste

Navn, firma, tlf., epost, adresse

Navn	Firma	Telefon	Epost
Anna Kim	Pexip AS		anna.kim@pexip.com
Knut Inge Hvidsten	Pexip AS		knut.hvidsten@pexip.com
Tomas Holt	NTNU	73559570	tomas.holt@ntnu.no
Fredrik Ruud	NTNU	93626455	fredrruu@stud.ntnu.no
Martin Almenningen	NTNU		

6.3 Avtaledokumenter

6.3.1 Arbeidskontrakt for bachelor-gruppen

Se vedlagt avtale.

6.3.2 3-partsavtale

Se vedlagt avtale.

B. Tidsplan

Non-English Text-to-Speech synthesis for video conferencing assistant

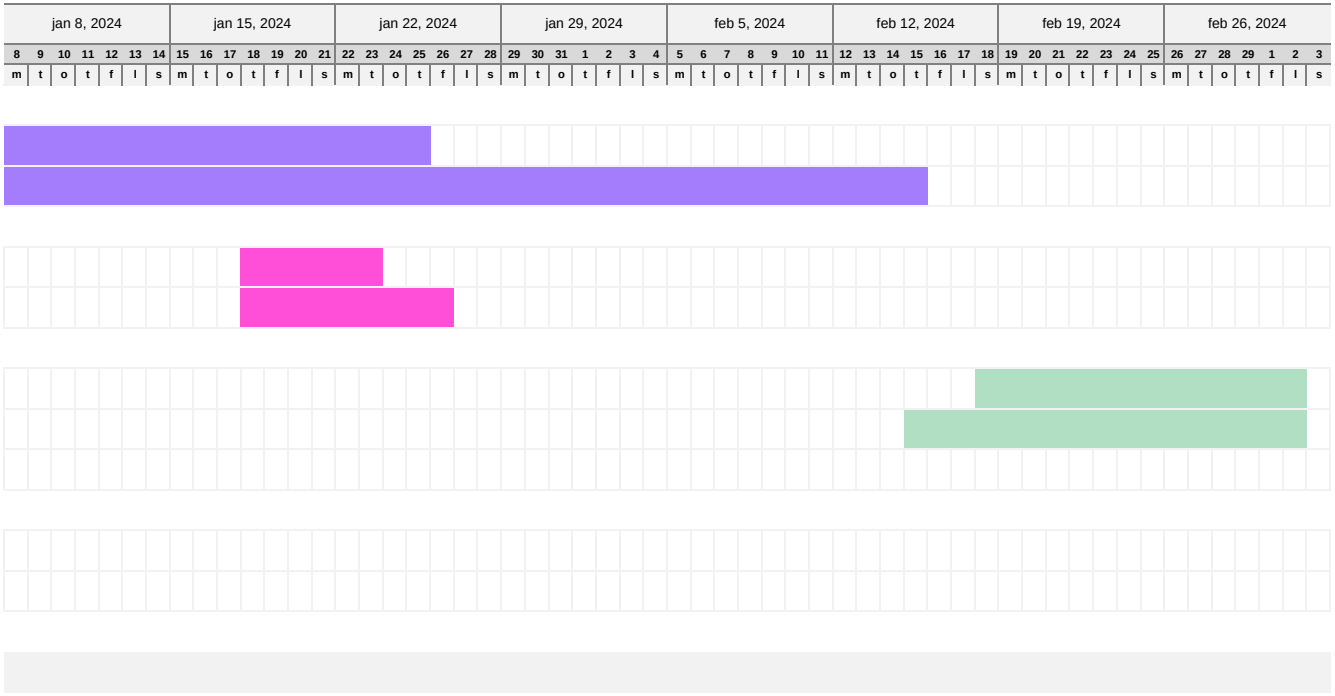
Pexip AS

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

Oppstart: man, 1.8.2024

Vis uke: 1

TASK	PROGRESS	START	END
Oppstart			
Definere mål		1.8.24	1.25.24
Lese seg opp på emnet		1.8.24	2.15.24
Planlegging			
Lage fremdriftsplan		1.18.24	1.23.24
Lage Forprosjektplan		1.18.24	1.26.24
Gjennomføring			
Poster		2.18.24	3.18.24
Utvikle modell		2.15.24	4.25.24
Testing og vurdering		3.26.24	5.10.24
Rapportering og evaluering			
Rapport		3.11.24	5.20.24
Presentasjon		5.20.24	5.25.24



Non-English Text-to-Speech synthesis for video conferencing assistant

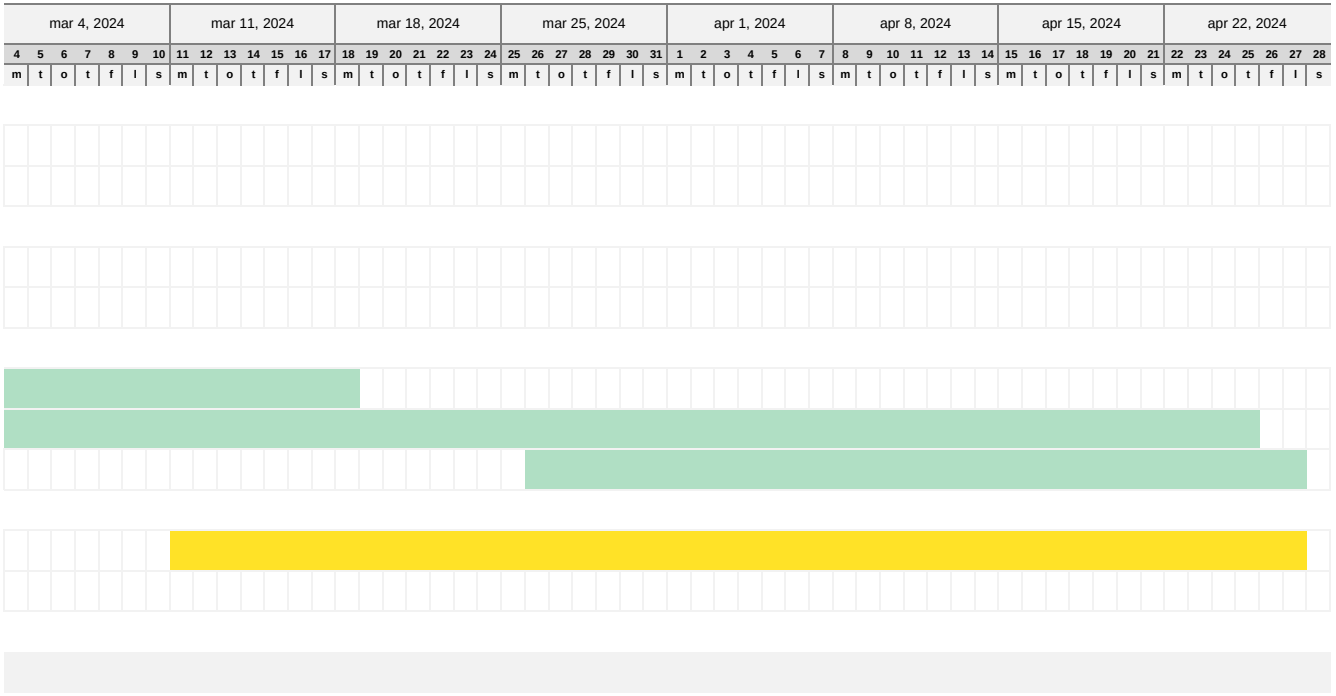
Pexip AS

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

Oppstart: man, 1.8.2024

Vis uke: 9

TASK	PROGRESS	START	END
Oppstart			
Definere mål		1.8.24	1.25.24
Lese seg opp på emnet		1.8.24	2.15.24
Planlegging			
Lage fremdriftsplan		1.18.24	1.23.24
Lage Forprosjektplan		1.18.24	1.26.24
Gjennomføring			
Poster		2.18.24	3.18.24
Utvikle modell		2.15.24	4.25.24
Testing og vurdering		3.26.24	5.10.24
Rapportering og evaluering			
Rapport		3.11.24	5.20.24
Presentasjon		5.20.24	5.25.24



Non-English Text-to-Speech synthesis for video conferencing assistant

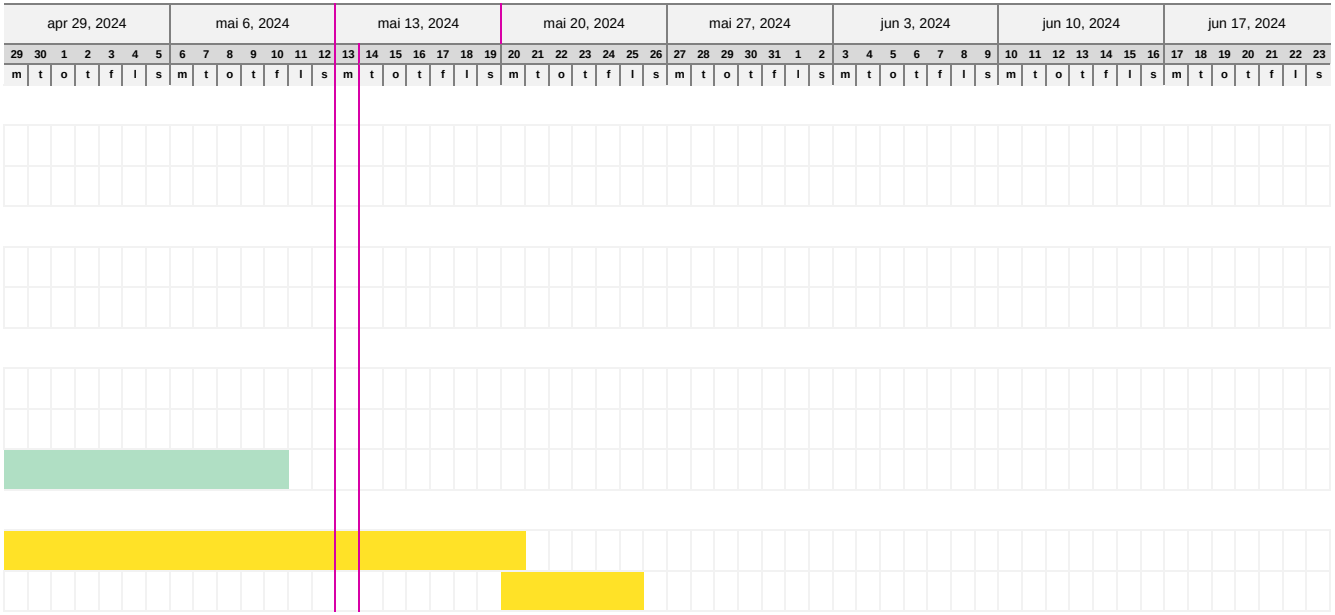
Pexip AS

SIMPLE GANTT CHART by Vertex42.com
https://www.vertex42.com/ExcelTemplates/simple-gantt-chart.html

Oppstart: man, 1.8.2024

Vis uke: 17

TASK	PROGRESS	START	END
Oppstart			
Definere mål		1.8.24	1.25.24
Lese seg opp på emnet		1.8.24	2.15.24
Planlegging			
Lage fremdriftsplan		1.18.24	1.23.24
Lage Forprosjektplan		1.18.24	1.26.24
Gjennomføring			
Poster		2.18.24	3.18.24
Utvikle modell		2.15.24	4.25.24
Testing og vurdering		3.26.24	5.10.24
Rapportering og evaluering			
Rapport		3.11.24	5.20.24
Presentasjon		5.20.24	5.25.24



C. Arbeidskontrakt

Arbeidskontrakt for Bacheloroppgave

Medlemmer: Martin Almenningen, Fredrik Bache Ruud

…….

Innledende tekst

Denne arbeidskontrakten bygger på et sett med typiske mål, oppgavefordelinger, prosedyrer og retningslinjer for interaksjoner for studentarbeider. Arbeidskontrakten er utfylt med *egne* fortolkninger av hva man mener med disse og hvordan man skal oppnå dette.

Mål

Effektmål

Hovedmålet er å utvikle en løsning som vil være særlig verdifull for brukere som kommuniserer på mindre utbredte språk, med et spesielt fokus på norsk. Oppgaven innebærer å utvikle en modell som kan lese opp norsk tekst på en naturlig, ikke-robotisk stemme.

Pexip planlegger å bruke prosjektet i forbindelse med videomøter der de har en modell allerede som oversetter og lager et sammendrag av et møtetranscript. Over ett langsiktig mål ønskes det at TTS- modellen blir iverksatt her sånn att sammendraget kan bli lest opp. Dette vil betydelig forbedre tilgjengeligheten og brukervennligheten for slike språkbrukere.

Modellen skal helst også kunne kjøres lokalt av brukere med egne maskiner. Dette hjelper til med å eliminere dyre serverkostnader, som mange andre moderne maskinlæringsmodeller ofte krever.

Dette prosjektet er ikke bare en mulighet til å bidra med vår ekspertise, men også en sjanse til å videreutvikle våre ferdigheter i et meget aktuelt og fremtidsrettet område.

Resultatmål

Spesifiserte Mål: Ved prosjektets slutt, skal teamet ha utviklet og implementert en TTS-modell som kan lese opp norsk tekst på en måte som er naturlig, klar og ikke-robotisk.

Målbare Kriterier: Suksesskriteriene for prosjektet inkluderer:

- En nøyaktighetsrate på minst 95% i tekst-til-tale konvertering.
- Brukertilfredshet målt gjennom en undersøkelse, med mål om minst 4 av 5 i gjennomsnittlig score.

Aksept og Realisme: Disse målene er utformet i samarbeid med alle prosjektmedlemmer. De er vurdert som oppnåelige innenfor prosjektets omfang og ressurser.

Tids- og Kostnadsrammer: Prosjektet skal fullføres innen en tidsramme på ca. fire måneder, uten ytterligere kostnader utover de allerede tildelte ressursene.

Roller og oppgavefordeling

Gruppen består av to medlemmer, og vi har derfor ingen leder.

Tittel	Hva det vil si	Ansvarlig
Dokumentansvarlig	Sørger for at dokumenter blir levert til riktig tid.	Fredrik
Referent	Skrive ned møtereferater til møtene	Martin
Møteinnkaller	Kaller inn til møter med arbeidsgiver og veileder.	Fredrik

Ved kode og dokumentasjon, skal arbeidet kvalitetssikres av personen som ikke har skrevet det.

Prosedyrer

A. *Møteinnkalling*

Medlemmene skal ha møter to ganger i uken, men kan ha det oftere om nødvendig. Møtene blir innkalt gjennom meldinger (Messenger).

B. *Varsling ved fravær eller andre hendelser*

Dersom man kommer for sent eller ikke kan møte, sier man i fra på Messenger så fort som mulig.

C. *Dokumenthåndtering*

Medlemmene jobber med delte dokumenter som blir lagret i OneDrive. Dokumentene blir diskutert under møtene slik at medlemmene alltid er oppdatert på hva som må gjøres og hva som er gjort.

D. *Innleveringer av gruppearbeider*

Medlemmene diskuterer fristene i møtene og dersom noe oppstår og det blir forsinkelser underveis, tar man det opp slik at det kan løses i god tid før fristen. Ideelt er arbeidet ferdig minst en dag i forkant slik at medlemmene kan revidere innholdet.

Interaksjon

A. *Oppmøte og forberedelse*

Medlemmene skal møte opp til tide. Tillater 10 minutters forsinkelse og eventuelt mer dersom vedkommende har en god unnskyldning. Medlemmene er forventet å være forberedt til de avtalte møtene.

B. *Tilstedeværelse og engasjement*

Det er tillatt å ta seg en pause eller gjøre noe annet under arbeidstid, så langt arbeidet som individet skal gjøre blir gjort og ikke distraherer de andre medlemmene.

C. *Hvordan støtte hverandre*

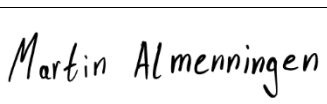
Medlemmene skal kommunisere misnøye og problemer med de andre medlemmene dersom det er noen. Det blir holdt en oppsummering på slutten av alle møtene der hvert medlem kan ta opp eventuelle problemer.

D. *Uenighet, avtalebrudd*

Ved uenighet prøver medlemmene å komme til et tilfredsstillende kompromiss. Dersom dette ikke lar seg gjøre tar medlemmene opp problemet med veileder og ser etter en løsning. Til, slutt hvis vi fremdeles ikke blir enige, bestemmer vinneren av stein, saks, papir.

Avtalebrudd blir diskutert i gruppen slik at det ikke oppstår flere ganger. Dersom et medlem bryter avtaler gjentagende ganger, blir det tatt opp med veileder, og dersom det fortsetter etter dette blir medlemmet kastet ut av gruppen.

Signatur

Fredrik Bache Ruud	
Marin Almenning	

D. Standardavtale batcheloroppgave

NTNU
Norwegian University of Science and Technology

Approved by the Pro-Rector for Education 10 December 2020

STANDARD AGREEMENT

on student works carried out in cooperation with an external organization

The agreement is mandatory for student works such as master's thesis, bachelor's thesis or project assignment (hereinafter referred to as works) at NTNU that are carried out in cooperation with an external organization.

Explanation of terms

Copyright
Is the right of the creator of a literary, scientific or artistic work to produce copies of the work and make it available to the public. A student thesis or paper is such a work.

Ownership of results
Means that whoever owns the results decides on these. The basic principle is that the student owns the results from their own student work. Students can also transfer their ownership to the external organization.

Right to use results
The owner of the results can give others a right to use the results – for example, the student gives NTNU and the external organization the right to use the results from the student work in their activities.

Project background
What the parties to the agreement bring with them into the project, that is what each party already owns or has rights to and which is used in the further development of the student's work. This may also be material to which third parties (who are not parties to the agreement) have rights.

Delayed publication (embargo)
Means that a work will not be available to the public until a certain period has passed; for example, publication will be delayed for three years. In this case, only the supervisor at NTNU, the examiners and the external organization will have access to the student work for the first three years after the student work has been submitted.

1. Contracting parties

The Norwegian University of Science and Technology (NTNU) Department: Department of Computer Science
Supervisor at NTNU: Tomas Holt email and telephone: tomas.holt@ntnu.no, 73559570
External organization: Pexip Email & phone: ole.andreas@pexip.com 916 46 103
Student: Fredrik Bache Ruud Date of birth: 25.04.1998
Student: Martin Nordli Almenningen Date of birth: 29.11.2001

The parties are responsible for clearing any intellectual property rights that the student, NTNU, the external organization or third party (which is not a party to the agreement) has to project background before use in connection with completion of the work. Ownership of project background must be set out in a separate annex to the agreement where this may be significant for the completion of the student work.

2. Execution of the work

The student is to complete: (Place an X)

A master's thesis	
A bachelor's thesis	X
A project assignment	
Another student work	

Start date: 11.01.2024
Completion date: 20.05.2024

The working title of the work is: Non-English Text-to-Speech synthesis for video conferencing assistant

The responsible supervisor at NTNU has the overarching academic responsibility for the design and approval of the project description and the student's learning.

3. Duties of the external organization

The external organization must provide a contact person who has the necessary expertise to provide the student with adequate guidance in collaboration with the supervisor at NTNU. The external contact person is specified in Section 1.

The purpose of the work is to carry out a student assignment. The work is performed as part of the programme of study. The student must not receive a salary or similar remuneration from the external organization for the student work. Expenses related to carrying out the work must be covered by the external organization. Examples of relevant expenses include travel, materials for building prototypes, purchasing of samples, tests in a laboratory, chemicals. The student must obtain clearance for coverage of expenses with the external organization in advance.

The external organization must cover the following expenses for carrying out the work:

Coverage of expenses for purposes other than those listed here is to be decided by the external organization during the work process.

4. The student's rights

Students hold the copyright to their works [1]. All results of the work, created by the student alone through their own efforts, is owned by the student with the limitations that follow from sections 5, 6 and 7 below. The right of ownership to the results is to be transferred to the external organization if Section 5 b is checked or in cases as specified in Section 6 (transfer in connection with patentable inventions).

In accordance with the Copyright Act, students always retain the moral rights to their own literary, scientific or artistic work, that is, the right to claim authorship (the right of attribution) and the right to object to any distortion or modification of a work (the right of integrity).

A student has the right to enter into a separate agreement with NTNU on publication of their work in NTNU's institutional repository on the Internet (NTNU Open). The student also has the right to publish the work or parts of it in other connections if no restrictions on the right to publish have been agreed on in this agreement; see Section 8.

[1] See Section 1 of the Norwegian Copyright Act of 15 June 2018 [Lov om opphavsrett til åndsverk]

5. Rights of the external organization

Where the work is based on or further develops materials and/or methods (project background) owned by the external organization, the project background is still owned by the external organization. If the student is to use results that include the external organization's project background, a prerequisite for this is that a separate agreement on this has been entered into between the student and the external organization.

Alternative a) (Place an X) General rule

X	The external organization is to have the right to use the results of the work

This means that the external organization must have the right to use the results of the work in its own activities. The right is non-exclusive.

Alternative B) (Place an X) Exception

X	The external organization is to have the right of ownership to the results of the task and the student's contribution to the external organization's project

Justification of the external organization's need to have ownership of the results transferred to it: If we do not own the results here we might end up in IP conflicts down the road

6. Remuneration for patentable inventions

If the student, in connection with carrying out the work, has achieved a patentable invention, either alone or together with others, the external organization can claim transfer of the right to the invention to itself. A prerequisite for this is that exploitation of the invention falls within the external organization's sphere of activity. If so, the student is entitled to reasonable remuneration. The remuneration is to be determined in accordance with Section 7 of the Employees' Inventions Act. The provisions on deadlines in Section 7 apply correspondingly.

7. NTNU's rights

The submitted files of the work, together with appendices, which are necessary for assessment and archival at NTNU belong to NTNU. NTNU receives a right, free of charge, to use the results of the work, including appendices to this, and can use them for teaching and research purposes with any restrictions as set out in Section 8.

8. Delayed publication (embargo)

The general rule is that student works must be available to the public.

Place an X

x	The work is to be available to the public.

In special cases, the parties may agree that all or part of the work will be subject to delayed publication for a maximum of three years. If the work is exempted from publication, it will only be available to the student, external organization and supervisor during this period. The assessment committee will have access to the work in connection with assessment. The student, supervisor and examiners have a duty of confidentiality regarding content that is exempt from publication.

The work is to be subject to delayed publication for (place an X if this applies):

Place an X		Specify date
	one year	
	two years	
	three years	

The need for delayed publication is justified on the following basis:

If, after the work is complete, the parties agree that delayed publication is not necessary, this can be changed. If so, this must be agreed in writing.

Appendices to the student work can be exempted for more than three years at the request of the external organization. NTNU (through the department) and the student must accept this if the external organization has objective grounds for requesting that one or more appendices be exempted. The external organization must send the request before the work is delivered.

The parts of the work that are not subject to delayed publication can be published in NTNU's institutional repository – see the last paragraph of Section 4. Even if the work is subject to delayed publication, the external organization must establish a basis for the student to use all or part of the work in connection with job applications as well as continuation in a master's or doctoral thesis.

9. *General provisions*

This agreement takes precedence over any other agreement(s) that have been or will be entered into by two of the parties mentioned above. If the student and the external organization are to enter into a confidentiality agreement regarding information of which the student becomes aware through the external organization, NTNU's standard template for confidentiality agreements can be used.

The external organization's own confidentiality agreement, or any confidentiality agreement that the external party has entered into in collaborative projects, can also be used provided that it does not include points in conflict with this agreement (on rights, publication, etc). However, if it emerges that there is a conflict, NTNU's standard contract on carrying out a student work must take precedence. Any agreement on confidentiality must be attached to this agreement.

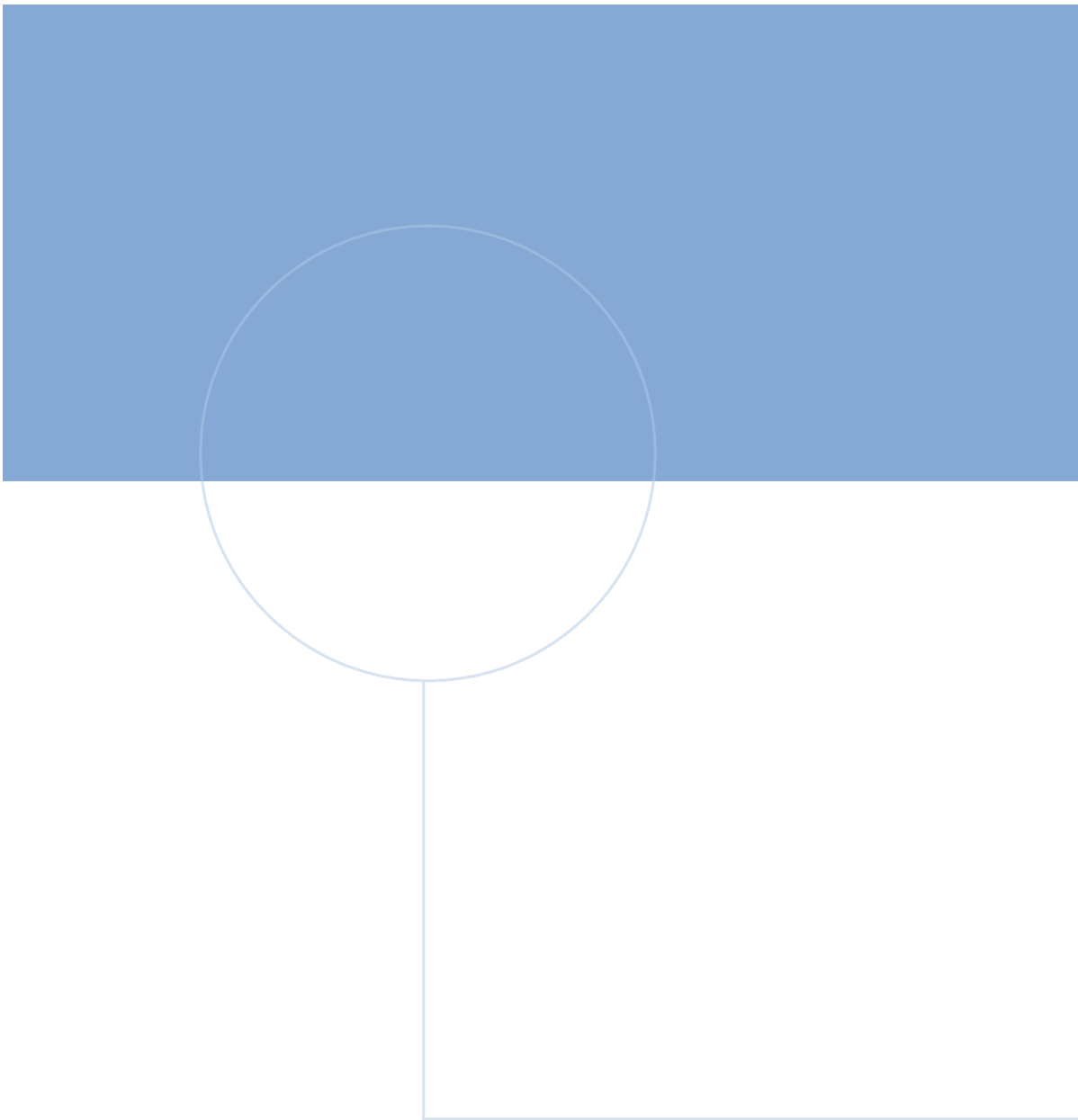
Should there be any dispute relating to this agreement, efforts must be made to resolve this by negotiations. If this does not lead to a solution, the parties agree to resolution of the dispute by arbitration in accordance with Norwegian law. Any such dispute is to be decided by the chief judge (sorenskriver) at the Sør-Trøndelag District Court or whoever he/she appoints.

This agreement is signed in four copies, where each party to this agreement is to keep one copy. The agreement comes into effect when it has been signed by NTNU, represented by the Head of Department.

Signatures:

for

Head of Department: Date: 02.04.24	
Supervisor at NTNU: Date: 02/2024	Tomas Holt
External organization: Date: 30/01/2024	
Student: Date: 01/02/2024	
Student: Date: 01/02/2024	Martin Almenningen



NTNU
Kunnskap for en bedre verden

]pexip[