Projektplan - Fagprojekt
Bachelor i Kunstig Intelligens og Data

Gustav G. Larsen (s180820), Lukas Leindals (s183920) & Peter Grønning (s183922)

June 2020

Figure 1: <3

Contents

1 Log book

The main purpose of the logbook is that it serves as a tool to keep track of the project and document project meetings.

1.1 Templates

Project Meetings

Questions
Reading, who and what
Implementation, who and what
Results, who and what
Decisions, who and what, what do you do alone, what do you do together

Supervisor Meetings

Presentation of results since last meeting
Action points for next week

1.2 Logs

NB: if nothing mentioned: everybody participated in the meeting.

Week 2

Date: *10-02-2020 to 17-02-2020*
Samarbejdskontrakt

- Produced with Gustav as referant
- Slack produced for professional messages: Join here
- Git repository created: Find it here
- Acknowledged by everyone - Maybe we need some slight adjustments
- To be signed.

Gant Chart

- In the making
- Skal Doing udvides?

Project Canvas

- Will be made at next meeting

Learning Objectives

- Will be made at next meeting

Next meeting: Tuesday 18-02-2020: *9:00 - 11:00 (12:00)*

Week 3

Date: *17-02-2020 to 24-02-2020*
We met wednesday for the lecture and afterwards a meeting with Morten Mørup. This gave us inspiration to continue the work on our project canvas and what project ideas to pursue.

Ideas for the project description

- Many - to - Many / Zero Shot
- Improve conversion with less target data - How little do we need to achieve fair conversion?
- Extract Speech features and styles - 'latent representation'
- Help improve conversion with little amount of target data - Select important data for conversion (Active learning?)
- Implement a real time solution for voice conversion.
- **Used Cases**
  - Repair Voices - Voice Impairment / Speech Enhancement
  - Create Deep Fakes - Karaoke
  - 'Reveal' Deep Fakes
  - Voices for animation movies
  - Voice Conversion across languages?

We decided on working with converting voices and see if we can optimise the training process to see how little take we need to succesfully convert a voice

Project canvas and Gantt chart

- Project canvas was worked on after having brain stormed ideas for the project description
- A lot of new actions was added to the trello board
- A time frame for the tasks was set

Learning Objectives
We formulated the learning objectives:

- Understand and use relevant terminology and theory regarding Deep Voice Conversion (DVC).
- Analyse and evaluate DVC models using appropriate statistical tools.
- Implement and improve DVC models.
- Discuss potential uses and misuses of DVC and ethical considerations regarding such technology.

Next week

- To do before next meeting (25-02-2020)
  - Read articles on the state of the art and possible models to use
  - Make projektbeskrivelse (will be done at this meeting)
  - Approve Gantt Chart
- Begin with the introduction
- Decide on methods
- Begin implementation of model
- Read more articles

Week 4

Date: *24-02-2020 to 02-02-2020*

Deciding on a project

- We discussed some articles about state of the art and the methods they used (GAN, AdaGAN, auto-encoding)
- We decided on working with converting accents, but ran into a concern about whether we could succeed or not, as we might run into trouble finding examples of code and data.
- We started the project description and made two descriptions one for general VC and one about converting accents

Formulated Project Description

- Together we formulated a project description and agreed upon scope and goals. (see project description)

Week 5

Date: *02-03-2020 to 09-03-2020*

Data collection

- We downloaded a large data set

ThinLinc

- playing around with thinlinc
- learned how to store the large data set on the remote computer and access on our local

Week 6

Date: *09-03-2020 to 16-03-2020*

Data

- Lukas had to work, while Peter and Gustav played around with the data set

Model

- starting to implement a model and see if we could reproduce others code

Meeting

- Peter and Lukas had a meeting with Morten, while Gustav was home working on the introduction and data part of the report

Adjusting to corona

- As we were not allowed to meet we set up a discord server that was used for future meetings and group work
- As our report was written in overleaf, the transition did not have much of an influence on the writing of the report

Week 7

Date: *16-03-2020 to 23-03-2020*

Mid-way report

- This week was mainly about working on the introduction, ethics, data and methods part of our report
- Peter worked a lot with Auto_VC and processing of audio signals
- Gustav worked a lot on the introduction, ethics and data parts
- Lukas worked a lot with the wavenet model

Week 8

Date: *23-03-2020 to 30-03-2020*

Feed back

- Tuesday we had a meeting where we sat down and gave feed back on the report "Training ASR models on synthetic speech as low-cost alternative to real data"
- Peter started implementing speaker identity encoder and GE2E loss

Who does what?

- Meeting on Wednesday
- Peter and Gustav will continue working on AutoVC
- Lukas will start looking at StarGAN
- The group will have status meetings every Saturday and Wednesday from now on and focus will be on trying to implement the model

Week 9

Date: *30-03-2020 to 06-04-2020*

- Speaker Identity encoder works
- t-SNE dimen

TODO

- Lukas looks into DTU HPC Cluster
- Lukas looks further into StarGAN
- Gustav will try to combine the AutoVC generator with Speaker Identity encoder
- Peter will try to implement a working loss function for the generator
- Peter will look a into t-SNE and write sections on speaker identity.

Will catch up Saturday 01-04-2020

Week 10

Date: *13-04-2020 to 20-04-2020*

HPC

- shell scripts was made to easily run the scripts via DTU's HPC clusters

Auto-VC

- The model scripts was adopted to run DTU's HPC clusters

Week 11

Date: *20-04-2020 to 27-04-2020*

AutoVC

- model seemed to produce an output, but learning curve looked suspicious

Week 12

Date: *27-04-2020 to 04-05-2020*

StarGAN

- Repo was downloaded
- The code was structured, so it followed the same structure as the AutoVC model wrt. the dat directory

AutoVC

- We discovered annealing rate is an important thing, which lead to a much prettier learning curve

Meeting with Morten

- Talked about reasons for weird learning rate

Week 13 + Exam period

Date: *04-05-2020 to 11-05-2020*

- We implemented the annealing rate that we discovered was crucial to the model last time, and it gave significant changes to the overall performance of the model.
- A few sections of the models section was fine written and made more cohesive

- we prepared for the exam period, meaning the project was set on hold due to the fact that we had to study for other courses.
- during the exam period: the Models were tested and developed a little further, nothing significant was achived.

Week 14 (first week of 3-week period)

Date: *04-06-2020 + 08-06-2020 to 12-06-2020*

04-06-2020

- First day of the last 3 weeks. We met up and discussed the overall plan for the next 3 weeks and compared it to our Gant diagram to check if we were up to date.
- We began to check all the viable possibilities for test data. It seemed hopeless to find anything legal, so we set up at meeting with some of our female class mates for sound files to test.

The following week:

- We found AmericanRhetoric.com, a giant speech bank with American speeches from across time. This gives us the possibility to test data legally from the selection of non copyrighted files.
- The danish state of ministry gives the new year speeches to download for free. We downloaded these and used the wav files to also test for danish speakers!
- We tested autoVC on the soundfiles and it turned out pretty good in some instances.
- A survey was created so we could evaluate our conversion with the general public. We used Shiney R for this since it seemed to be the easiest best survey tool we could access with the needs to play sound files.
- The general Survey/Experiment was designed so we knew what to test for and what to leave out. However we still needed StarGAN to train on the soundfiles which takes a lot of time.. will most likely first be available early next week.

Week 15 (second week of 3-week period)

Date: *15-06-2020 to 19-06-2020*

- We met almost everyday to sit together and work as a team
- The survey was sent out to the public for data gathering to get some results
- The methods section had some changes to contain all the underlying theory of some methods as well

- the Experiment section was started with adding information sheet and experimental design with illustrations.
- Scope and introduction was revisited to contain the true scope we came to have in this project

Week 16 (third week of 3-week period)

Date: *22-06-2020 to 24-06-2020*

- The last writing face was initiated and results and discussion were started on and finished.
- A grand, long read through the report was begun to make sure every nook and cranny was understandable and made sense.
- The report was handed in.

2 Project description

2.1 Motivation

We see a tendency in voice conversion being more normalised in our everyday. Both on social media, such as Snapchat where filters give you a wacky new voice, and perhaps soon also in more business oriented cases like call centers trying to make their voice sound different so they may seem more reliable from a customers point of view[1]. We also see voice conversion spread into the news as Deep-Fakes, manipulating and fooling people[2]. In a world where most people still see their voice as unique as their fingerprint, we find it interesting to explore the world of converting voices in a believable way. This raises the question of who can we trust and how easy is it to fool these people?

2.2 Scope

The scope of this paper revolves around investigating state-of-the-art Voice Conversion(VC) models and will test them on the basis of training data used for natural voiceconversion, as well as try to convert voices of famous speakers convincingly. In this paper the two state-of-the-art VC models, AutoVC and StarGAN, will be implementedand tested. Furthermore, the results obtained will be compared to the results of theoriginal papers. Conversion speech will be synthesised using different vocoders.The success criteria of this project would be to implement the models in such a way that the final product would sound convincing to the human ear by testing our conversionswith an experiment/survey.
Our success criteria is that the final product will be able to succeed a Turing Test (a human not being able to distinguish the machine output from a real human being). And hopefully we will be able to implement the models in such a way that we can reduce the time and data needed for training and keep the quality of the conversion high.

Research Questions

- Q1: How do state-of-the-art models alter voices and how can these be implemented?
- Q2: How well do the state-of-the-art methods perform when converting the voices of famous speakers?
- Q3: What ethical and lawful restrictions exists in the field of deep fakes?

[1] Justin Calderon: Inside the secret world of accent training, BBC

[2] Catherine Stupp: Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, The Wall Street Journal

3 Learning goals

- Understand and use relevant terminology and theory regarding Deep Voice Conversion (DVC).
- Analyse and evaluate DVC models using appropriate statistical tools.
- Implement and improve DVC models.
- Discuss potential uses and misuses of DVC and ethical considerations regarding such technology.

4 Samarbejdskontrakt

4.1 Møder

Tid og sted

Et møde er hvor man samles om fagprojekt. Det kan både være til at gøre status, men også at arbejde videre på projektet.

- Som udgangspunkt hver onsdag på DTU, hvor der er plads.
- Vi har intentioner om at svømme hver onsdag om eftermiddagen.
- Mandage lyder som en dag hvor folk tager ud på DTU, så der kan møder nemt ske.
- 6-10 timer om ugen ca. til møder og individuelt arbejde i 13-ugers.

Afbud

Hvis man er bliver nød til at melde afbud, skal det gøres senest dagen inden møde.
Hvis man på dagen bliver forhindret i at møde, meld ud så tidligt som muligt.

Referat

Referat implementeres i logbogen som punktform af de mest vigtige informationer fra mødet.

Roller

Roller udvælges ved random sampling uden tilbagelægning SEED = 'dagens dato - ddmmyyyy'

- Referent
- Ordstyrer
- Koordinator (Koordinere næste møde)

Beslutninger

Beslutninger tages med følgende workflow:

1. Enighed - Hele gruppen skal være enig om en beslutning
2. Kompromis - Kan enighed ikke opnås forøgs dette at indgå et kompromis.
3. Afstemning - Hvis et kompromis ikke kan findes, tages beslutningen ved afstemning.

4.2 Kommunikation i gruppen

Praktisk

Såsom mødetider og -steder og lignende over messenger chat.

Faglig

Fagligt: Relevant materiale, links, filer, repositories over Slack.

4.3 Indsats og forventninger

Flextid

I alt antager vi, at vi får lavet lige meget, men det kan variere efter ansvarspunkter, og hvor vi er i forløbet, hvor meget vi hver især kan lave, og derfor behøver alle ikke at lave lige meget altid.

Forventninger

- Det forventes, at alle leverer lige meget til projektet.
- Det forventes, at alle arbejder mod et godt produkt med god kvalitet, men at man har det fedt med projektet og dets indhold kommer i første række.
- Det forventes, at alle kan forstå og forklare om de forskellige dele af emnet/rapporten.
- Deadlines overholdes. Hvis man ikke kan nå at aflevere til tiden, så skal det udmeldes i god tid, og udskudt tidsplan skal godkendes af hele gruppen.
- Det forventes, at alle sætter sig ind i relevant materiale, som deles via slack.
- Det forventes, at logbogen holdes up-to-date - Ugens referant er ansvarlig for dette. Hele gruppen skal godkende ugens logbog.
- Det forventes, at der er god stemning under arbejdsprocessen, og man altid kan udtrykke sin holdning til en aktivt lyttende gruppe.
- Det forventes, at vi alle er modtagelige overfor konstruktiv kritik.
- Brug af Git:
  - Det forventes at, GitHub bruges som den primære platform til udarbejdning af projektet samt versionskontrol.
  - Det forventes at der arbejdes på seperate branches, når nye dele udforskes og udvikles. På denne måde vil master branch altid være fuldt fungerende.
  - Man holder sig til den respektive branch.
  - Det forventes, at alle 'puller' inden arbejde og 'pusher' efter endt arbejde.

  - Alt relevant materiale ligges op i GitHub, så alle altid kan tilgå alt materiale.
  - Det forventes at commit beskeder er relevante og beskrivende for hvad der committes.

Ejerskab

Jeg, Lukas, vil bidrage til forventningerne med:

- Git og kodning
- Forsøge at hjælpe med overblik over hvor langt vi er i processen
- Læse op på relevant materiale
- Animationer

Jeg, Gustav, vil bidrage til forventningerne med:

- Skrive og læse op på Etik i forhold til projektet
- Git og kodning
- Underholde gruppen
- Holde struktur i rapporten

Jeg, Peter, vil bidrage til forventningerne med:

- Git og kodning
- Informationssøgning
- Holde styr på arbejdsprocessen

4.4 Problemer

Ved problemer forstås:

- Interne uenigheder (socialt eller fagligt)
- Gruppemedlem overholder ikke aftaler
- Fejlkommunikation
- Metodeproblemer - hvilken skal vi benytte?
- Alvorlig sygdom

Problemløsning:

- Ved hvert møde afsættes tid til at udarbejde problemer.
- Hvis vi ikke kan løse et problem selv, går vi til vejleder.

4.5 Underskrifter

Lukas Leindals (s183920): 26-02-2020

Peter Grønning (s183922): 26-02-2020

Gustav Gamst Larsen (s180820): 26-02-2020

5 Feedback givet til os

- Flot rapport
- Der kan læres noget fra vores rapport
- Vi har styr på meget teori og har skrevet en god del teori
- Kunne være rart med et afsnit der er en lille opremsning af metoderne og lige skabe et godt overblik over alt den teori der benyttes. Det er svært at holde styr på det og få lagt en nem rød tråd ud fra en.
- Fjerner "vi"
- Hvordan benytter vi selv de teorier vi gennemgår? hvordan bliver de implementeret i det afsluttende produkt?
- Turing test - implementere turing test i metode afsnittet? - mean opinon score (skala på 1-5, hvor meget minder her om en ægte stemme?)
- You are the target audience
- en én sætnings forklaring for hvad en deepfake er (måske som en fodnote)
- Hvad kan man gøre for at standse sådan en teknologi hvis den kommer i de forkerte hænder? (måske i diskussionen)
- kant og kantisme / dentological etik
- LSTM - husk kilde
- Gode til at sørge for at læseren forstår alt teorien, godt beskrevet!

6 Project Canvas

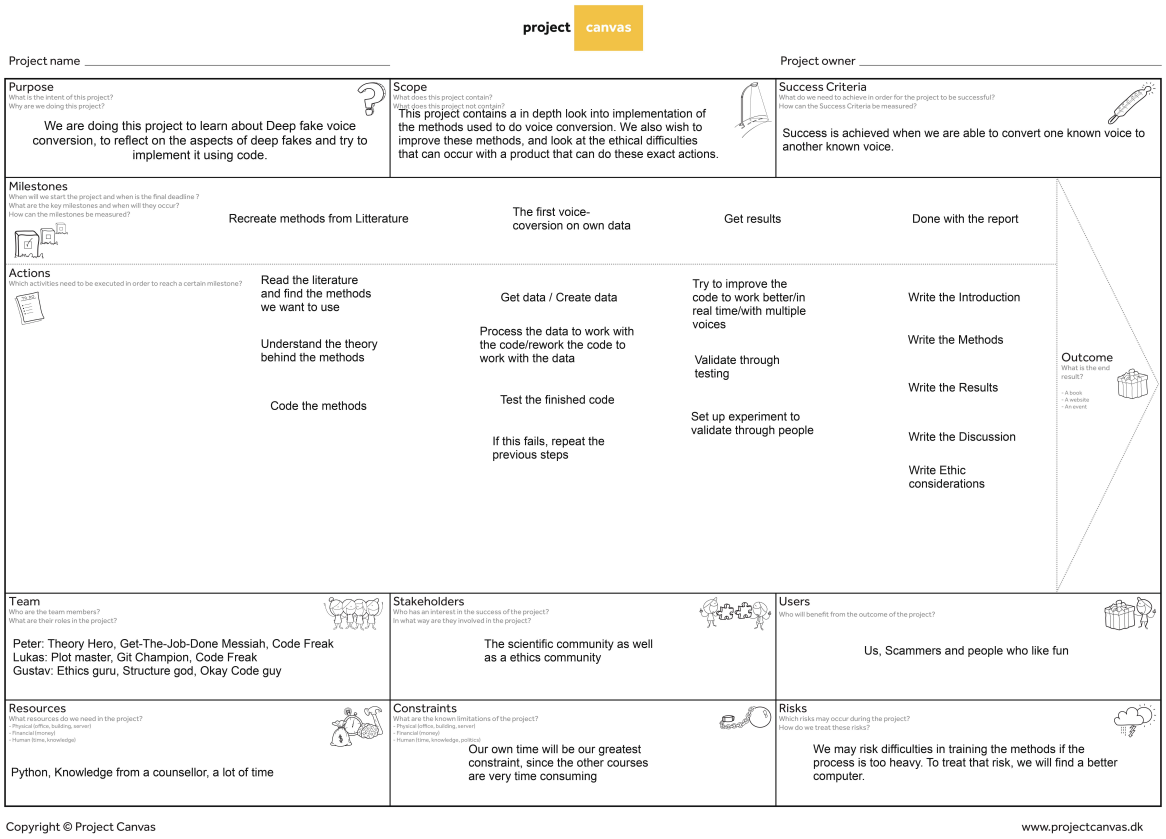
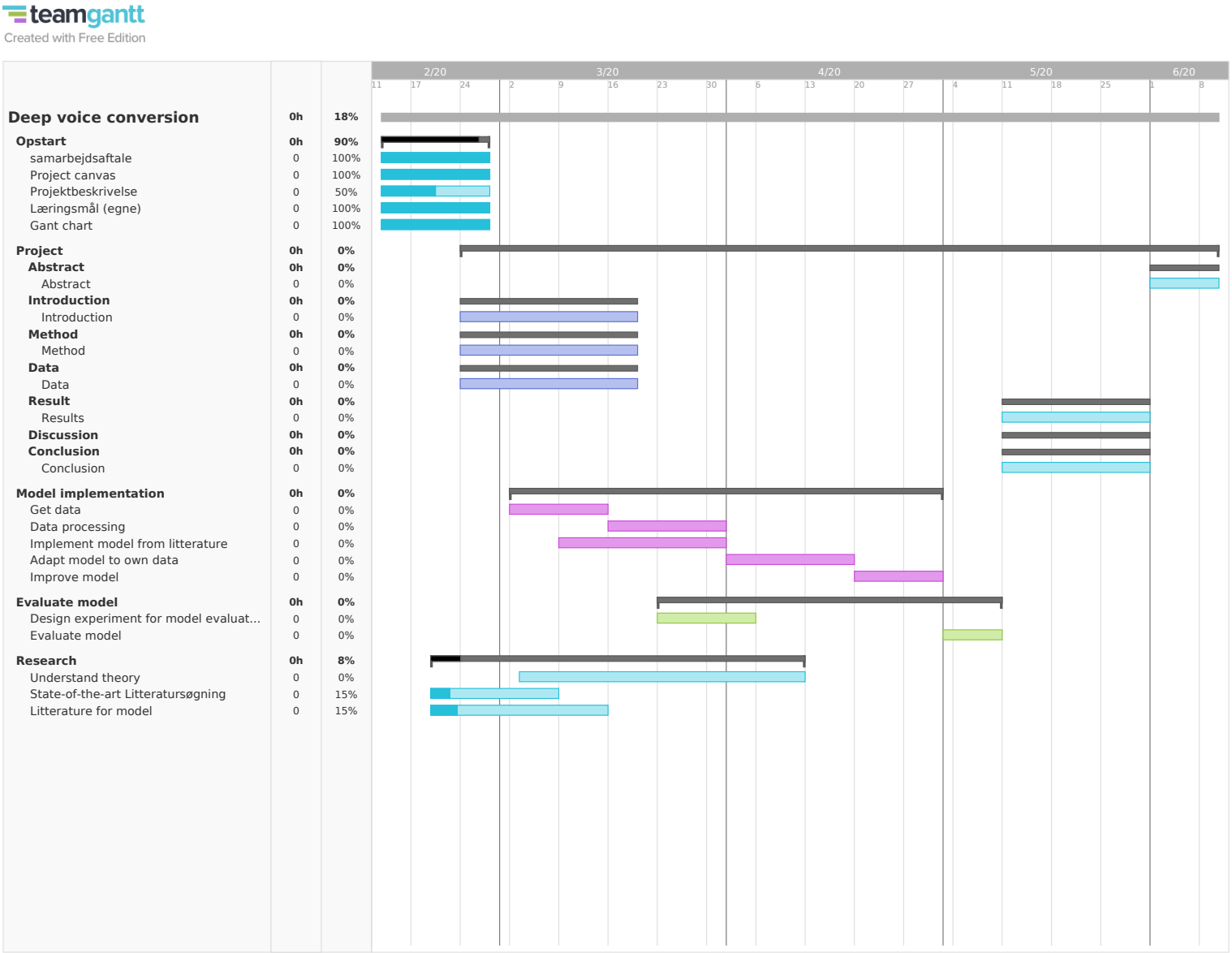
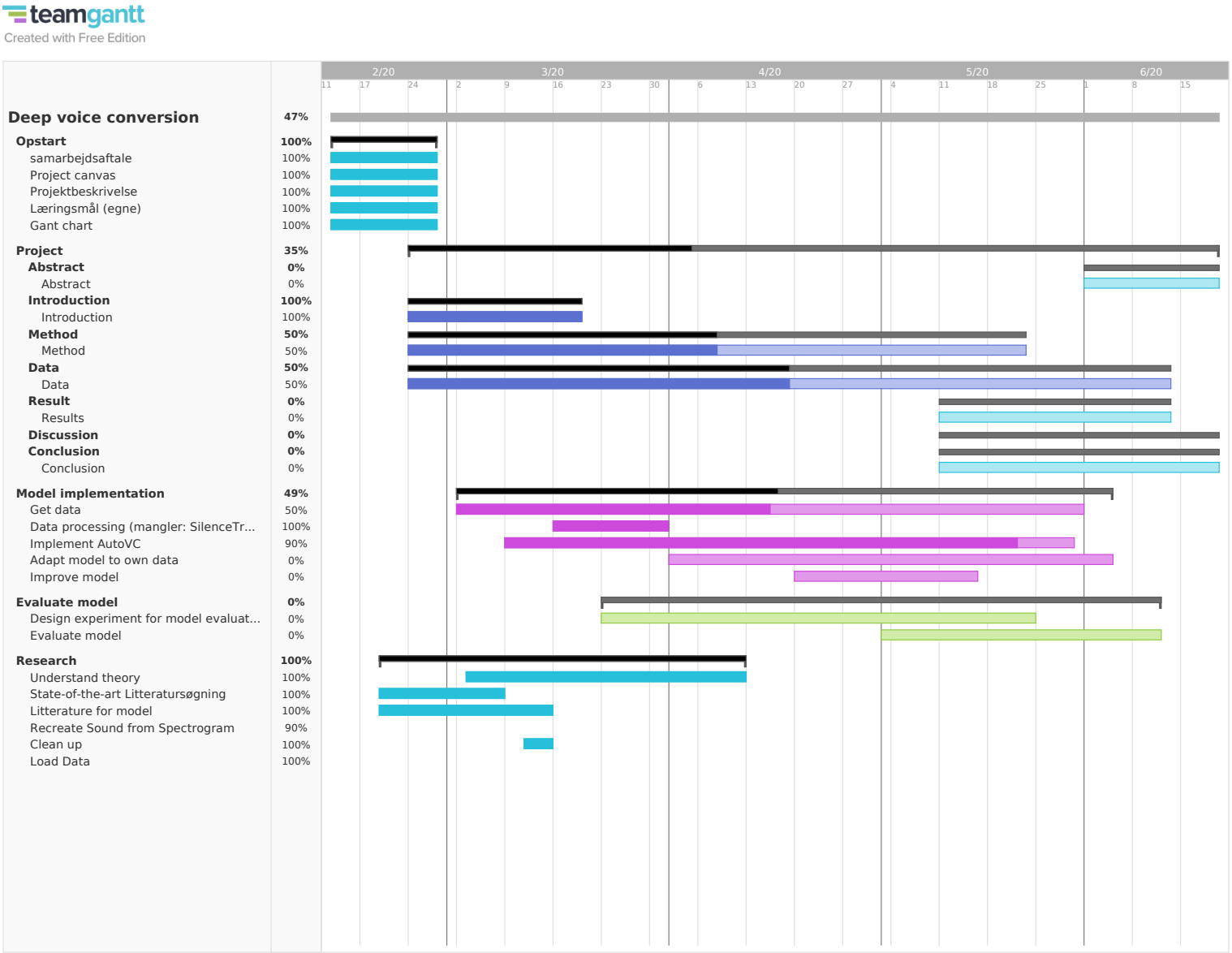


Figure 2: Vores originale Project Canvas skrevet de første 2 uger af fagprojektet. Vi har siden da ændret bl.a. scope da vi indså arbejdsmængden i projektet.

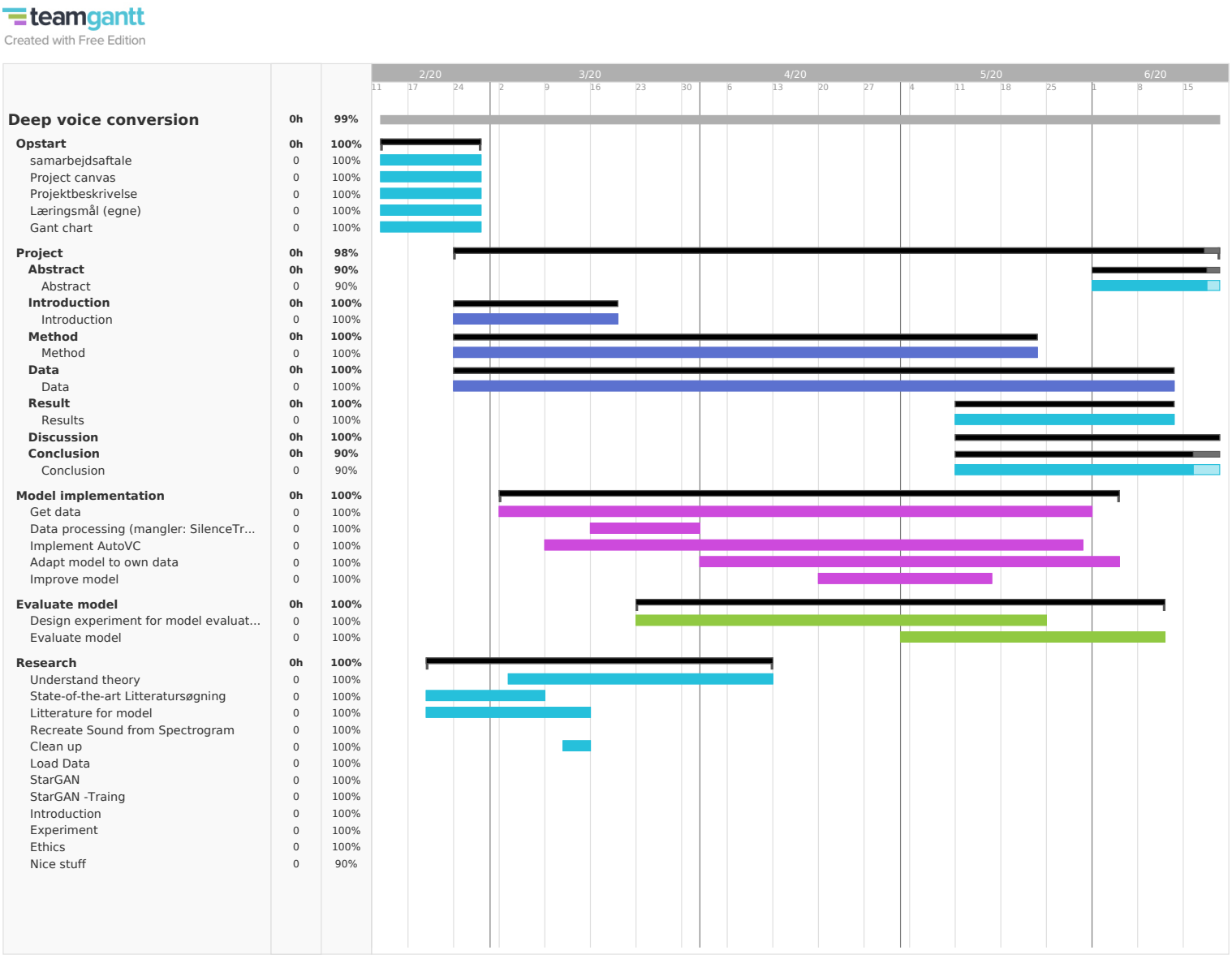
7 Project Revision



7.1 Gantt lige før 3 ugers perioden



7.2 Gantt på dagen vi afleverede fagprojektet



7.3 Autoevaluring

Opdateret Fagprojekt Plan - Voice Conversion

Gustav Larsen - 180820
Lukas Leindals - s183920
Peter Grønning - s183922

May 2020

1 Revuderet Tidsplan

- Sværere at implementere metoderne end forventet
- Hvad gør vi med data?
- Main fokus er bare at få voice conversion til at fungere
- AutoVC er nogenlunde succesfuld, men mangler lidt
- StarGAN er påbegyndt som en anden metode
- Metode afsnittet skal skrives færdigt/muligvis cuttes i her og der

Oprindeligt forventede vi at have to state-of-the-art VC modeller AutoVC og StarGAN færdigt implementerede på nuværende tidspunkt. Denne forventning var baseret på, hvad vi troede omfanget af source-code var. Det viste sig eksempelvis, at store dele af AutoVC ikke var nemt tilgængeligt, herunder et Speaker Encoder modul og en træningsfunktion, som hhv. måtte findes andetsteds fra og implementeres selv. Det har med ført et uventet ekstra arbejde med PyTorch, hvilket har været en stejl læringskurve, idet vi ikke har arbejdet rigtigt med sådanne Deep Learning moduler tidligere i vores uddannelse. Desuden har specifikationerne for data håndteringen (lydsignal til mel spektrogram) været uklare, hvilket har ført til en lang og ikke færdig afsluttet parametertuning, så modellen kan spille sammen med vocoderen anvendt i AutoVC.

Ved siden af overraskende umedgørlige modeller kommer den tunge teori, som de bygger på. Meget af den underliggende teori er nyt og vanskeligt stof, som har krævet meget tid at sætte sig ind i. Vi må også tilkendegive, at den forståelse og forklaringsgrad af teorien, vores vejleder ønsker, ligger over vores tidlige forventninger. Det har medført et møjsommeligt arbejde, som har været værdifuld for vores forståelse, men samtidig har gjort, at flere deadlines måtte overskrides.

Grundet diverse copyright love etiske årsager har det også været svært at skaffe test data, hvilket vi stadig kæmper med.

Vores optimistiske forventninger om at kunne forbedre conversion i form af at reducere tid og data i processen har vi måtte skyde ned. Vi besidder ikke de rette forudsætninger til dette, så vi ændrer fokus til bare at få det implementeret i stil med litteraturen og gøre det brugbarte på det datasæt, vi nu måtte finde.

Det er lykkedes os, at implementere AutoVC og en dertilhørende trænings-funktion, og det virker lovende. Vi mangler dog, at få de rette spektrogram specifikationer, før vi kan bruge AutoVC's vocoder. StarGAN er undervejs og det næste skridt er at teste, hvorvidt det er rigtigt implementeret.

8 Feedback givet

Feedback for the report: "Training ASR models on synthetic speech as low-cost alternative to real data"

March 2020

1 Feedback

In general a well written project so far, great job! The problem is well defined and incorporated in great manner in an engaging introduction, which gives a good perspective on why the problem is chosen and why it is important. The concrete and limiting scope makes it clear, what the project is about and what you wish to investigate. The research questions are a bit hidden away and simple (yes/no - questions). Highlighting these would improve the introduction and considering a reformulation like "How does the ASR model perform..." would heighten the overall quality of the project outcome.
The use of figures is deliberate and helps with the understanding of concepts e.g. figure 3.2 carries section 3.0.2.3. Do not be shy with the use of figures, they are really helpful the ones you have, so whenever a section is 'figure-less' you miss them as reader.
The text is written in a way, which makes the rather difficult topics easy to understand. You are not only good with words but you manage to structure the project in a logical way and presenting the most important first. Use your writing skills to explain some of the many terms e.g Fast Fourier Transformation, CNN, ReLU, WaveNet etc. which otherwise are left unexplained. This would make you look more confident with the methods and increase the reproducibility of the project.

All in all the report is coming along really nicely. Your way of visualising and formulating methods and concepts gives the report a great quality as well as the aesthetically pleasing design and layout. It was kind of difficult to find a lot of criticism but the ones we did stood out. We believe that if you were to implement the feedback then the finished product would be fantastic since you are already travelling along an amazing path for victory and glory.

1

2 Questions for the feed back meeting

1. How does your research questions lead into a discussion in the later part of the report?
2. Who is your target audience and how well do they know the topic you are writing about?
3. How come only some sections have illustrations and not all?
4. What is the difference between log mel spectrogram and Mel Frequency Cepstrum Coeffiecients?
5. In what way will you use/implement data augmentation?
6. How do you intend to evaluate and validate the model?
7. How can you show of your work live?