Degree Project in Computer Science

Second cycle, 30 credits

Enhancing Norwegian Text-to-Speech: Developing a Proof-of-Concept by Applying an Iterative Model Training Approach

GARD ÅCKERSTRØM AASNESS

Enhancing Norwegian Text-to-Speech: Developing a Proof-of-Concept by Applying an Iterative Model Training Approach

GARD ÅCKERSTRØM AASNESS

Date: July 1, 2024

Supervisors: Ahmad Al-Shishtawy, Syed Zohaib Hassan, Pål Halvorsen
Examiner: Jonas Beskow
School of Electrical Engineering and Computer Science
Host company: SimulaMet
Swedish title: Förbättring av norsk text-till-tal: Utveckling av en Proof-of-Concept-modell genom att tillämpa en iterativ modellträningsmetod

Abstract

Text-to-speech (TTS) technology converts written text into synthesized speech. Developing a TTS system for low-resource languages such as Norwegian poses significant challenges due to the limited availability of high-quality, diverse datasets and open-source models. This thesis addresses the problem by developing a proof-of-concept (PoC) TTS model for Norwegian, focusing on adult speech as a precursor to developing a child speech TTS model in the future.

In this research, we iteratively trained and evaluated four TTS models using various datasets, including a multi-speaker dataset and two single-speaker datasets. A PoC adult speech TTS model was developed as a foundation for future transfer learning to create child speech TTS models, which are crucial for tools and applications used by children, providing an appropriate voice for their interactions. The models were assessed using both objective metrics, specifically Word Error Rate (WER), and subjective metrics, specifically Mean Opinion Score (MOS), to identify their strengths and weaknesses.

The main findings show that combining datasets enhances model performance, as demonstrated by Model 4, which achieved the lowest WER of 14.95% and the highest MOS with scores of 3.96 for intelligibility and 3.14 for naturalness. Additionally, it was found that a larger volume of data is crucial for training intelligible and natural TTS models, but starting with imperfect data can still yield significant results, even with as little as two hours of training data. These insights pave the way for future advancements in TTS technology, ultimately contributing to the creation of high-quality synthetic speech for various applications, including the development of child speech TTS models for scenarios such as police interview training.

Keywords

Text-to-speech, Proof-of-concept, Low-resource language, Norwegian, Adult speech, Child speech, Matcha-TTS, Mean opinion score, Word error rate, Data-driven iterative approach

Abstract

Text-to-speech (TTS) teknologin omvandlar skriven text till syntetiskt tal. Att utveckla ett TTS-system för språk med begränsade resurser, såsom norska, innebär betydande utmaningar på grund av den begränsade tillgången på högkvalitativa och varierade dataset samt öppen källkodsmodeller. Denna avhandling adresserar problemet genom att utveckla en proof-of-concept (PoC) TTS-modell för norska, med fokus på vuxental som en föregångare till att utveckla en TTS-modell för barnröst.

I denna forskning har vi iterativt tränat och utvärderat flera TTS-modeller med hjälp av olika dataset, inklusive ett dataset med flera talare och två dataset med enskilda talare. En PoC TTS-modell för vuxental utvecklades som en grund för framtida transfer learning för att skapa TTS-modeller för barnröst, vilket är avgörande för tillämpningar som realistiska barnavatarer som används i polisintervjuträning för att bekämpa barnmisshandel. Modellerna utvärderades med både objektiva mått, specifikt Word Error Rate (WER), och subjektiva mått, specifikt Mean Opinion Score (MOS), för att identifiera deras styrkor och svagheter.

De viktigaste resultaten visar att kombinationen av dataset förbättrar modellens prestanda, vilket demonstrerades av Model 4, som uppnådde den lägsta WER på 14,95% och den högsta MOS med betyg på 3,96 för begriplighet och 3,14 för naturlighet jämfört med enskilda modeller. Dessutom visade det sig att en större datavolym är avgörande för att träna förståeliga och naturliga TTS-modeller, men att börja med ofullkomliga data kan ändå ge betydande resultat, även med så lite som två timmars träningsdata. Dessa insikter banar väg för framtida framsteg inom TTS-teknologin, vilket slutligen bidrar till skapandet av högkvalitativt syntetiskt tal för olika tillämpningar, inklusive utvecklingen av TTS-modeller för barnröst för scenarier som polisintervjuträning.

Nyckelord

Text-till-tal, Konceptbevis, Resurssvagt språk, Norska, Vuxental, Barntal, Matcha-TTS, Genomsnittligt omdömesbetyg, Ordfelprocent, Datadriven iterativ metod

Acknowledgments

The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

I want to thank everyone at SimulaMet who helped make this thesis possible, with special thanks to Syed Zohaib Hassan, my external supervisor. He gave me the chance to work with him, scoped the assignment, and guided me through the development and writing process. His advice on what to do, how to prioritize, and his feedback on the report were invaluable. I also want to acknowledge Pål Halvorsen, chief research scientist, for his feedback on the work and report, and Tore H. Larsen, who helped me set up the eX3 and overcome challenges related to it.

A big thanks to Ahmad Al-Shishtawy, my internal supervisor from KTH, who supported me during the thesis search, provided feedback on the work, and helped me navigate the various processes involved in completing a master's degree. Thanks as well to Jonas Beskow, my examiner, for his valuable discussions, suggestions, and advice on some of the challenges I faced, which made the research complete.

Lastly, thanks to all the participants of the user surveys. Their contributions were essential for gathering the data needed to compare the models, a critical part of this thesis.

Thanks again to everyone who supported and contributed to this work.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Aritficial Intelligence.
ASR	Automatic Speech Recognition.
AUC	Area Under the ROC Curve.
CFM	Control Flow Matching.
CPS	Child Protective Services.
CPU	Central Processing Unit.
GDPR	General Data Protection Regulation.
GPS	Global Positioning System.
GPU	Graphics Processing Unit.
HMM	Hidden Markov Model.
Hz	Hertz.
ITU	International Telecommunication Union.
MAS	Monotonic Alignment Search.
ML	Machine Learning.
MOS	Mean Opinion Score.
MSE	Mean Squared Error.
NPSC	Norwegian Parliamentary Speech Corpus.
OT	Optimal-transport.
OT-CFM	Optimal-transport Control Flow Matching.
PoC	Proof-of-concept.

RMSE	Root Mean Squared Error.
ROC	Receiver Operating Characteristic.
SOTA	State-of-the-art.
TTS	Text-to-speech.
WER	Word Error Rate.
WSPSR	Web-scale Supervised Pretraining for Speech Recognition.

Chapter 1

Introduction

This section aims to give the reader an introduction to the thesis by providing the necessary information and context to understand what the thesis is about, its scope and facilitate the rest of the thesis. Specifically, the section describes the background of the thesis, the specific problem that the thesis addresses, the objectives, the ethical considerations, the scope and delimitations, and the structure of the report.

1.1 Background

Text-to-speech (TTS) refers to the process of transforming written text into synthetic speech, offering solutions to various societal and medical challenges. From aiding visually impaired individuals through voice assistants to enhancing user experiences in products like audiobooks and GPS systems, TTS has become ubiquitous in modern technology. Despite its widespread applications, the development of TTS systems faces notable challenges, particularly regarding the availability of diverse and high-quality speech datasets. There are very few available datasets for low-resource languages and child speech in particular. This scarcity hinders research in these areas and contributes to global issues, such as services and tools designed for children that use adult voices instead of age-appropriate ones. Additionally, independently creating these datasets is especially challenging due to the specific requirements and characteristics necessary for TTS datasets.

The motivation behind this research specifically is police interview training on children. Child abuse and maltreatment are a huge problem worldwide, leaving children with lifelong consequences and suffering, and worst case, even death [1]. Child sexual abuse significantly impacts global

health, requiring intervention from law enforcement and Child Protective Services (CPS) to protect children [2]. In cases of abuse, the children themselves often serve as crucial witnesses, yet the lack of corroborative evidence combined with no signs of physical abuse underscores the importance of their testimony in forensic interviews [3, 4]. Conducting these interviews requires understanding children's development and employing supportive questioning styles. However, numerous international studies conducted across various countries have highlighted the prevalence of inadequate investigative interviews as a significant concern [5].

To combat this problem of child abuse, a team at a Norwegian research center called SimulaMet is investigating the possibility of making a realistic virtual child avatar that police officers can practice their interview skills on. The goal is that police officers will be able to conduct better interviews in real-life scenarios by having a better arena to practice [6, 7].

Norwegian can be considered a low-resource language due to its limited resources and relatively small population of almost 5.5 million. As previously mentioned, child speech datasets are particularly scarce for low-resource languages. A common approach to addressing this challenge is transfer learning, where knowledge from one domain is applied to another, reducing the need for large datasets for the target domain. For creating a child speech TTS model, an adult speech model can serve as a base, learning basic intonation, pitch, flow, and other speech characteristics from adult speech before being fine-tuned with smaller child speech datasets to resemble child speech.

1.2 Problem Statement

Unfortunately, there are neither any available child speech nor adult speech TTS datasets in Norwegian, omitting transfer learning as an option. Consequently, this thesis will focus on creating a Norwegian adult speech TTS model. This research will serve as a foundational step toward developing a child speech TTS model for future applications, such as police training. However, the thesis will still research the development of a child speech TTS model as well, suggesting approaches to overcome challenges for future research.

1.3 Research Questions

Since there is no open-source Norwegian adult speech TTS model available, the goal of the thesis is to develop a proof-of-concept (PoC), which in turn highlights the challenges and opportunities regarding Norwegian TTS model development. To effectively reach the goal, two research questions have been formulated to ensure targeted work:

1. What type of speech data is most suitable for TTS development in a low-resource language?

2. How natural and intelligible can the synthetic speech generated by the PoC TTS model be?

Research question 1 will be answered by creating three different datasets based on speech gathered from various sources that will be used to train four different models, followed by a performance comparison between these models. Based on the models' performances, a conclusion can be drawn regarding which data type was most suitable, while also considering the effort required to create the datasets.

Research question 2 requires scores that represent the intelligibility and naturalness of the synthesized speech. Both subjective and objective metrics will be used. To score both intelligibility and naturalness subjectively, user surveys will be conducted, gathering the opinions of several participants and using them to calculate the mean opinion score (MOS). Additionally, intelligibility will also be scored objectively by using an automatic speech recognition (ASR) model to calculate the word error rate (WER), which represents how well it was able to understand the synthesized speech.

1.4 Research Methodology

To develop the TTS model, an iterative data-driven approach, also known as empirical training, was adopted. This approach leverages datasets to train models over several iterations, ensuring they learn the intricate patterns of natural speech from real-world examples. The empirical nature of this method allows the model to generalize well to unseen data, enhancing its robustness and overall performance.

For the evaluation of the TTS model, mixed methods were employed, combining both qualitative and quantitative assessments. Specifically, a user survey was conducted for qualitative evaluation, where human listeners

rated the naturalness and intelligibility of the synthesized speech, allowing a MOS to be calculated. This subjective evaluation provides direct insights into the perceived quality of the speech, capturing nuances that may not be reflected in quantitative metrics. In addition, an intelligibility assessment using an ASR model was performed for quantitative evaluation. This method involves transcribing the synthetic speech with an ASR model and comparing the transcriptions to the original text, calculating the WER to quantify intelligibility.

The choice of these methods is motivated by the lack of other quantitative assessment techniques due to the inherently subjective nature of speech. Additionally, combining qualitative and quantitative methods ensures a comprehensive assessment of the TTS model. While MOS captures human perceptions of speech quality, the ASR-based intelligibility assessment provides objective, measurable data, which combined creates a good balance and ensures reliable results.

1.5 Scope and Limitations

This thesis aims to develop a PoC mainly focusing on synthesizing natural and intelligible speech for adult speech rather than on inclusiveness in terms of synthesizing speech for different genders, dialects, and other factors.

While the research will explore the available data and approaches for creating TTS models for Norwegian child speech, the actual development of a child TTS model is beyond the scope of this work. Developing a child TTS model presents significantly greater challenges compared to adult speech TTS, primarily due to the scarcity of high-quality, annotated datasets, or even just available ”found” data sources, which is practically any data that was made for another purpose, such as podcasts or YouTube videos. Additionally, the complexities involved in accurately modeling child speech characteristics are a big challenge. The inclusion of information about child TTS in this thesis serves to provide context and background relevant to the original problem addressed by this research: creating a realistic child avatar for police interview training. By exploring the state of child TTS, this thesis aims to highlight the potential pathways and challenges in this area, thus setting the stage for future work. Additionally, the absence of an available open-source Norwegian adult speech TTS model further motivates the decision to implement a PoC for this purpose. This serves as a valuable stepping stone rather than directly developing a PoC for a child speech TTS model.

1.6 Ethical Considerations

The ethical considerations discussed in this section are constrained to the scope of this thesis, not to TTS development in general. When developing an adult TTS model, several ethical considerations must be taken into account to ensure responsible and fair development and use of the technology. Firstly, it is generally crucial to obtain informed consent from all participants whose voices are recorded specifically for usage in the dataset. However, in this thesis, most of the data used is publicly available, which indirectly means that the speaker has agreed to the possibility of someone using it. The only voice recorded is my own voice, which I give myself consent to use. Additionally, the purpose of this thesis is strictly academic. Development motivated by a commercial purpose would most likely set even stricter requirements.

Moreover, while it is generally important to address issues of bias and discrimination when training TTS models, by using diverse datasets that represent various dialects, genders, ages, and socio-economic backgrounds to avoid perpetuating existing biases and discrimination, the scope of this thesis is focused solely on creating a voice that is as natural and intelligible as possible. Therefore, considerations of inclusivity are not the primary concern in this work.

Lastly, the ethical implications of the model's deployment should be considered. Developers and users of TTS models should ensure that the technology is used in a manner that benefits society and does not contribute to harm. This includes avoiding uses that could lead to misinformation, harassment, or other malicious activities.

1.7 Structure of the Thesis

The structure of the report is as follows:

- Chapter 2 provides the necessary theory to understand the rest of the thesis, including information about TTS systems, TTS dataset characteristics, child speech characteristics among other topics.
- Chapter 3 is a literature review of related work, including recent advancements and the state-of-the-art (SOTA), and TTS for low-resource languages and child speech.
- Chapter 4 describes all necessary information to conduct the research, such as the data that was used and how it was collected, which

experiments was conducted, what testbed was used to conduct the experiment and how the final model was evaluated.

- Chapter 5 presents the results from the evaluation of the model.
- Chapter 6 interprets the results and discusses factors that potentially contributed to bias that might have impacted the results.
- Chapter 7 summarizes the work and the most important findings and provides some suggestions for future work.

Chapter 2

Theory

The theory section will provide the reader with all the background knowledge necessary to understand the rest of the thesis. Everything that is mentioned is related to the problem this thesis aims to investigate. The theory will consist of the following topics in order:

- Explaining the common development process of a TTS model and describing single-speaker and multi-speaker models.
- Deep learning, neural TTS architectures, and Matcha-TTS, which is the newly released deep learning based TTS architecture this thesis uses.
- Dataset characteristics for TTS development in general and for low-resource languages specifically.
- Objective and Subjective TTS model evaluation.
- Explaining child speech characteristics, focusing on what differentiates it from adult speech.
- Transfer learning, which is a technique often used in low-resource settings.

2.1 TTS Model Development

As mentioned in the introduction, TTS is the process of generating synthetic speech from text. Developing a TTS model often includes the same main steps. To begin with, as in any task centered around training a model, getting high-quality data is one of the most important steps. The output generated by the

model is not going to be any better than the input the model has trained on. If there is not any high-quality data available, there are several pre-processing measures that can be taken to improve the quality, such as data augmentation and noise removal. Two important terms when developing a TTS model are pre-trained model and fine-tuning. A pre-trained model refers to a model that has already been trained on data and can be used to satisfy a generic task. Fine-tuning refers to the process of tweaking the pre-trained model to a specific task by training it on data specific to that task. Most architectures also include pre-trained models, which for example have been trained on English speech data. If there is no already pre-trained model, you can create one yourself by training the model on more general data in the beginning. In the case of a Norwegian child TTS, a model can be pre-trained on Norwegian adult speech to learn the model general behavior before fine-tuning it to Norwegian child speech to learn the model specific behavior. This example highlights the importance of initially developing an adult speech TTS model to achieve a child speech TTS model later, which is what this thesis aims to do in the form of a PoC. When a model has finished training, it is important to evaluate it to observe how well it performs. That is done by calculating various metrics and often comparing them to a baseline model's performance. A TTS system is usually evaluated by its naturalness and intelligibility.

2.1.1 Single-speaker vs Multi-speaker

The development of a TTS model can be designed for either single-speaker or multi-speaker synthesis. A single-speaker model synthesizes speech from only one speaker, requiring a dataset with speech data solely from that speaker for training. In contrast, a multi-speaker model can synthesize speech from multiple speakers, requiring a dataset that includes speech data from at least two different speakers. The inclusion of multiple speakers in the dataset allows the model to generalize better and learn diverse speech patterns. However, it remains crucial to have sufficient data from each individual speaker to ensure the model can accurately reproduce the unique speech characteristics of each speaker during synthesis.

Single-speaker TTS models are simpler to design, implement, and train, as they focus on capturing the characteristics of one voice. This often results in higher quality and more natural-sounding speech for that specific speaker, with lower data requirements since only one individual's speech needs to be modeled. However, these models lack flexibility and can only generate speech in a single voice, making them unsuitable for applications that require multiple

voices.

On the other hand, multi-speaker TTS models offer greater versatility and flexibility, as they can generate speech in various voices, making them ideal for applications requiring multiple speakers, such as audiobooks or diverse voice assistants. By incorporating multiple speakers, these models can generalize better and learn a wider range of speech patterns. However, they are more complex to design and train, requiring larger and more diverse datasets.

2.2 Deep Learning

Deep learning is a subset of machine learning (ML), which involves algorithms and models inspired by the structure and function of the human brain, particularly neural networks, to learn from large amounts of data. As shown in Figure 1, neural networks are composed of multiple layers of interconnected neuron nodes. The first layer is the input layer and contains the input neurons, which consist of the data for the neural network. The number of layers can vary between models, and they can be both visible and hidden. The network processes the data by having each layer process the data from the previous neuron and perform transformations on it. The layers extract increasingly abstract features of the input data, and the final output layer produces the output of the model, such as a prediction, a classification, or a sequence of text or speech. Therefore, the more layers a neural network has, the more complex data it can work with.

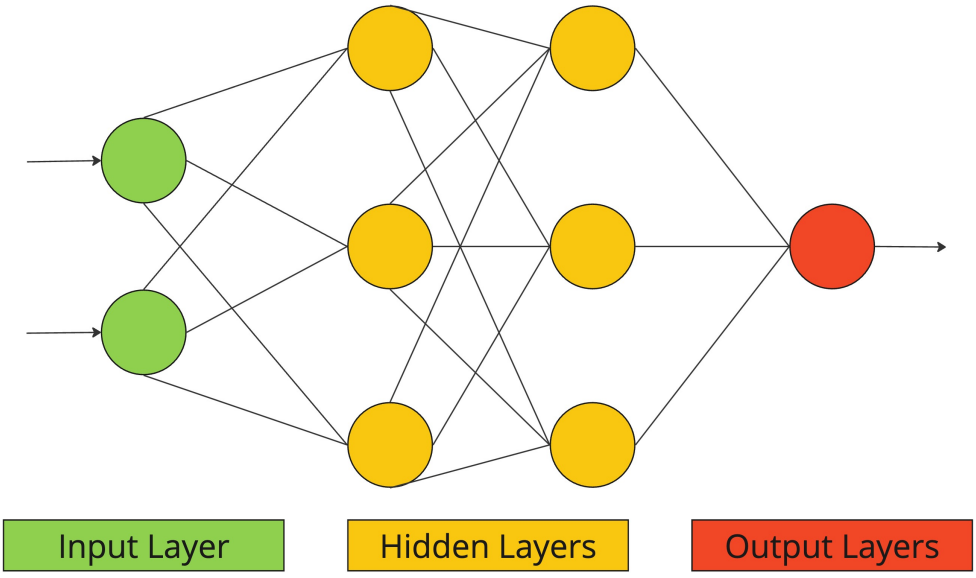

Figure 1: Architecture of a neural network with an input layer, three hidden layers, and an output layer. Adapted from Sportfire: https://www.spotfire.com/glossary/what-is-a-neural-network

.

2.2.1 Neural TTS Architecture

This section will discuss the architecture of neural TTS exclusively, seeing how most modern models are deep learning-based. The architecture of a neural TTS system can vary depending on the focus of the TTS system. However, as shown in Figure 2, a general neural TTS system consists of three different components, namely text analysis, acoustic model, and vocoder. Text analysis takes a sequence of text as input and converts it into phoneme or grapheme features. These features are then passed into the acoustic model, which converts them into acoustic features. Acoustic features can be viewed as an abstract representation of the speech waveform, and the most commonly used acoustic feature in neural TTS is the mel spectrogram. Finally, the vocoder converts the acoustic features into waveforms, which is the final format of speech and the output of the neural TTS model [8].

Figure 2: General components in a neural TTS architecture, adapted from Tan et al. [8].

2.2.2 Matcha-TTS

In recent years, deep learning and neural networks have become the preferred approaches for developing high-quality text-to-speech (TTS) architectures and models due to their efficient learning capabilities and ability to capture complex structures. This section will specifically delve into the architecture of Matcha-TTS [9], a new deep learning-based TTS architecture introduced in 2023, which will be utilized for training the models in this thesis.

Matcha-TTS introduces two significant innovations that mitigate the trade-off between speed and quality: an improved encoder-decoder architecture that reduces memory consumption and accelerates evaluation and the use of optimal-transport conditional flow matching (OT-CFM) for training the model. An encoder-decoder architecture consists of an encoder that processes input data and converts it into an internal representation known as a context vector, and a decoder that utilizes this context vector to produce the desired output. These architectures are particularly suited for TTS tasks due to their flexibility in handling inputs and outputs of varying lengths and their memory efficiency by compressing information into a context vector. Moreover, they are well suited for low-resource settings by using data efficiently and enabling alignment-free training, which helps the model generalize well on limited data [9].

As previously mentioned, Matcha-TTS employs OT-CFM as its training method, unlike many models, particularly older ones, which rely on score matching. OT-CFM is designed to align the probability distributions of data by transporting samples from one distribution to another while maintaining certain constraints. Optimal Transport (OT) addresses the minimization problem of finding the most efficient way to move mass between two distributions. Conditional Flow Matching (CFM) involves transforming samples from one distribution to another through specific transformations. This approach is particularly beneficial for TTS, as it enables the model to effectively learn to convert text features, such as phonemes, into speech features, such as spectrograms, while considering variables like speaker identity. The use of OT-CFM allows Matcha-TTS to train high-quality speech models more efficiently and in fewer steps compared to score matching [9]. In short, score matching is a method for estimating model parameters by matching the gradients of the log-probability of the data to the gradients of the model, especially useful when calculating the full probability is difficult.

Additional design choices make the architecture well-suited for low-resource languages and capable of synthesizing natural-sounding speech

quickly. Firstly, the model learns to speak from scratch without external alignments, enabling it to establish the relationship between text and speech directly from the data without needing pre-existing pairings. This reduces the demand for extensive data, which is beneficial when data availability is limited. Secondly, the model is probabilistic, meaning it uses probabilities to generate speech, allowing for slight variations in the output for a given input. This variability can enhance the naturalness of the speech. Lastly, the model is non-autoregressive, generating all parts of the speech simultaneously instead of one word at a time, unlike autoregressive models that produce speech token by token based on the previous token. This non-autoregressive approach significantly improves synthesis speed [9].

Matcha-TTS enables multi-speaker training through its flexible encoder-decoder architecture and the integration of speaker embeddings. These embeddings serve as vector representations of different speakers' voices, conditioning the model to generate speech in the desired voice. Additionally, OT-CFM allows the model to learn mappings from text to acoustic features while incorporating speaker-specific characteristics. This architecture, combined with efficient memory usage and mechanisms like Monotonic Alignment Search (MAS), ensures accurate alignment and duration prediction for multiple speakers, enabling high-quality multi-speaker TTS synthesis [9]. Ensuring a balanced representation of different speakers in the training data can be challenging, and there is a risk of regression to the mean, where the unique characteristics of individual speakers might be averaged out, resulting in less distinctive voices. However, the probabilistic nature of Matcha-TTS helps mitigate this risk by capturing the full distribution of speech features for each speaker and maintaining their unique characteristics. In contrast, non-probabilistic models often struggle with this issue, as they risk averaging out speaker-specific features, leading to less personalized and distinctive speech outputs. Despite these challenges, multi-speaker TTS models provide a scalable solution for generating diverse and personalized speech outputs.

2.3 Dataset

A general rule for developing an AI model is that a model's performance is not going to be any better than the quality of the data it is trained on. If a model is trained on bad data, the output will be just as bad. However, different AI models require datasets with different characteristics specifically suited for that task. This section will discuss the various characteristics a dataset dedicated to TTS training requires, followed by some measures that can be taken for low-

resource languages where no dataset with the required characteristics might exist. This knowledge is important to understand the challenges that will be faced when creating a high-quality speech dataset and the effort it takes.

2.3.1 TTS Dataset Characteristics

To understand the characteristics of the TTS dataset, it is essential to understand the goal of a TTS model. The goal of a TTS model is to transform the written text into clear and natural-sounding synthetic speech. To achieve that, it is important to reduce the amount of noise and other disturbances in the speech data used to train the model as much as possible. If the speech data the model trains on is noisy, it teaches the model that the synthetic speech should also be noisy. In contrast, an ASR model has the complete opposite goal, where it wants to recognize speech in noisy environments and transform it into written text. Therefore, it is not uncommon to inject noise into the audio of an ASR speech dataset to make it more robust. This comparison highlights the difference in dataset characteristics between different tasks.

Other characteristics TTS datasets have are low values for mean pitch, standard deviation of energy, speaking rate, and level of articulation [10]. Therefore, the creation of high-quality TTS datasets often requires professional equipment, professional readers, and a recording studio, which is both expensive and time-consuming. This is one of the main reasons high-quality TTS datasets are limited compared to high-quality ASR datasets. Professional readers are typically instructed to speak with as little variation as possible to reach the consistently low values as previously mentioned. Therefore, the sentences read are usually planned and not spontaneous as they normally are for ASR recordings [10].

Furthermore, a sign of a high-quality TTS dataset is a normalized distribution of audio duration between 1 and 20 seconds, ensuring enough short and long utterances so the model will handle synthesizing speech well regardless of the text length.[1]

A TTS dataset can be either single-speaker or multi-speaker, indicating whether the speech is produced by one speaker or multiple speakers. Research suggests that using a large amount of speech from a single speaker is often preferred [11, 12], likely because it better captures speech patterns, resulting in more accurate voice reproduction. However, other studies argue that utilizing available multi-speaker data is equally effective, if not superior, to using

[1] https://docs.coqui.ai/en/latest/what_makes_a_good_dataset.html

extensive single-speaker data [13]. In conclusion, it shows that all possibilities should be considered when choosing what data to use, especially in a low-resource language where the options might be limited.

2.3.2 Dataset for Low-resource Languages

In situations involving low-resource languages, such as Norwegian, the options for selecting speech data are severely constrained due to the scarcity of available resources. Consequently, one must work with whatever data is accessible, employing a range of strategies to optimize outcomes from this limited foundation. However, some important discoveries can help train a good model even for low-resource languages. To begin with, for a TTS developer experimenting with low-resource languages, who is starting with limited or no data at all, it is recommended to find radio broadcast news or audiobooks, as they exist in most languages, and produce relatively intelligible and natural voices [14]. That is because they share many similarities to TTS data compared to other speech data sources, such as having a lot of speech from a single speaker and the other characteristics mentioned in subsection 2.3.1 [14]. In her research, Cooper states that as little as five minutes of high-quality data can improve the model [14]. After obtaining a small amount of high-quality data, if available, it is possible to perform data augmentation on it, which synthetically increases the size of data by tweaking various factors such as speaking rate and pitch.

As for ASR corpora, it is more difficult to generate intelligible voices. However, it is possible to use it with a fair amount of noise cleanup. As most languages typically have more ASR data than TTS data, this can be used as low-quality data to improve the model if only a minimal amount of high-quality data exists [14]. In addition, being selective with the utterance selection when using ASR data, only using the data that satisfies the criteria of a standard deviation of f0, fast speaking rate, and hypo-articulation, can help produce more intelligible voices [15].

2.3.3 Child speech datasets

Creating child speech datasets, whether using found data like YouTube videos or other methods, faces significant challenges. Firstly, there is a limited number of videos with children as main speakers, and the utterances in these videos are often short. Additionally, background noise and music in these videos degrade audio quality. The lack of proper transcriptions and

annotations further complicates the use of such data for training. Collecting high-quality child speech data typically requires controlled environments, proper recording equipment, and the recruitment of child speakers, which introduces challenges such as data protection laws, the involvement of parents or guardians for consent, and the need to manage children's shorter attention spans and lower levels of focus [16].

2.4 TTS Model Evaluation

Unlike other AI tasks, including ASR and predictive models such as regression- and classification tasks, which can evaluate the performance of models using objective measures such as WER, area under the ROC curve (AUC), root mean squared error (RMSE) and mean squared error (MSE), evaluating a TTS model is not as straightforward. That is because the factors that determine the quality of speech, such as naturalness and intelligibility, vary between listeners. That means that what one person thinks sounds natural, another person might think sounds unnatural. Therefore, the quality of human speech is mainly a subjective evaluation, as it is based on human perception.

The most common way to subjectively evaluate a TTS model is by having several people listen to chosen audio snippets while rating the speech based on its naturalness and intelligibility. After gathering scores from all the participants, MOS is calculated, which represents the opinion of the average listener. Figure 3 shows the formula for calculating MOS, where R refers to the ratings and N refers to the number of participants.

$$\text{MOS} = \frac{R_1 + R_2 + \ldots + R_N}{N}$$

Figure 3: Formula to calculate MOS

Despite being the most common practice for evaluating a TTS model, there are still a couple of shortcomings and considerations that are important to be aware of. To begin with, the quality of synthetic speech should be evaluated based on the expectations of listeners in regard to the specific application context and domain in which it will operate [17]. For example, a voice that is good for reading audiobooks in a quiet environment might not be the optimal voice for a noisy environment. Second, a study performing an analysis of the MOS test methodology has discovered that researchers use

the term inconsistently and often underreport the details of their evaluation methods [18]. Common mistakes are having different ranking scales, having bad communication with the participants about what they should base their evaluation on, and excluding essential information such as the number of participants and so on. Consequently, this leaves a lot of room for misinterpretation, both for the participants evaluating the models and for people reading the research. It presents the question regarding what the evaluation really is [18]. This highlights the importance of performing a good evaluation process, both in terms of the actual testing and the documentation in the report.

Despite MOS being the most common metric for TTS models, using an ASR model to transcribe synthetic speech and calculate the WER can also be done to get a more objective assessment of intelligibility. Figure 4 shows the formula to calculate the WER score, which is the sum of errors (S+I+D) divided by the total number of words (N). Errors consist of substitutions, insertions, and deletions. Substitution signifies that a wrong word has been recorded, insertion means that an additional word has been recorded by mistake, and deletion means that a word is missing and therefore has not been recorded.

$$\text{WER} = \frac{S + I + D}{N}$$

Figure 4: Formula to calculate WER score

2.5 Child Speech Characteristics

Child speech is different from adult speech due to several reasons, such as higher pitch, longer duration, different formant frequencies, less developed vocabulary, and different pronunciation. It is important to know the differences to understand the challenges associated with developing synthetic child speech. All the differences between the speech highlight how different they are, and why it is so important to generate synthetic child speech as well. This section will be based on a research paper by Lee et al. that performs an analysis of children's speech [19].

2.5.1 Duration

To begin with, vowel duration for both genders shares the same peak at age five, showing a significant difference from older children until age seven. Children's tendency to overshoot or in some cases undershoot vowel duration may suggest that the dynamic range of vowel duration is larger for children than for adults. Because they more often than not overshoot the vowel duration, the rate of speech is slower for children. Vowel duration variability measures the inconsistency in vowel timing across different instances of speech, while vowel duration magnitude indicates the average duration of vowels within a speech sample. Analysis of sentence duration suggests that both magnitude and variability reach adult levels around age eleven or twelve. This implies that coarticulation skills may influence both duration and variability [19].

2.5.2 Pitch

Second, children have a higher pitch which gradually decreases the older they get. The average pitch of a six-year-old child is about 275 Hz, showing no significant difference between genders. The pitch gradually decreases for both genders until they are about twelve years old, when the female pitch starts to plateau at 225 Hz, compared to the male pitch that hits a steep development around puberty and flattens out at around 125 Hz. Therefore, the average pitch of the speech between children and adults can vary up to 150 Hz depending on age and gender [19].

2.5.3 Formant Frequencies

Formant frequencies are peaks in the sound waves produced when we speak. They occur because of the specific shape and configuration of our vocal tract. These peaks enhance certain frequencies in speech sounds, making vowels and consonants sound distinct. Formant frequencies play a crucial role in how we perceive and produce different sounds in spoken language, and are therefore very relevant to synthetic speech generation. There are differences in formant frequencies between children and adults, with male and female patterns starting to differentiate around age ten or eleven and becoming fully distinguishable around age fifteen. Male formant frequencies decrease faster, reaching the adult range around age fifteen, while for females, it's around age fourteen [19].

2.6 Transfer Learning

Transfer learning is a machine learning technique where a model trained for a specific task is reused or adapted for another task. According to Pan and Yang in their article ”A Survey on Transfer Learning”, a definition of transfer learning is:

”Given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t, transfer learning aims to help improve the learning of the target predictive function $f_t(\cdot)$ in D_t using the knowledge in D_s and T_s, where $D_s \neq D_t$, or $T_s \neq T_t$.”[20]

A large amount of data is required to develop a deep learning model from scratch, which might not be available. This technique is often used when there are limited data or other resources, as the model can learn from the knowledge gained from training on a larger related dataset. One of the most common ways the transfer learning approach is used is by fine-tuning a pre-trained model, as previously mentioned in section 2.1.

Chapter 3

Related Work

3.1 Recent Advancements and SOTA

TTS has changed a lot over the years, with more powerful techniques leading to better and more natural-sounding synthetic speech in recent years. In 2022, Kaur and Singh provided a clear and concise review of the conventional and contemporary approaches used in TTS synthesis [21]. During the 2000s leading up until recent years, the statistical parametric approach such as the Hidden Markov Model (HMM) was widely used to generate synthetic speech [22]. The use of HMM helped develop speech synthesis as it is very flexible in modeling various aspects of it, such as phonetic units, prosody, and duration, which makes it adaptable to different speaker styles and languages. In addition, the theoretical background is well established and provides a good foundation for further research on HMM and synthetic speech. However, the use of HMM showed weakness in other important aspects of a synthetic speech model, such as relying on large amounts of data for training, which proved to be difficult for low-resource languages. Additionally, the synthetic speech lacks naturalness, which is an important factor in evaluating a TTS model. These weaknesses, among others, opened up room for improvement [22, 23].

In more recent years, various deep learning techniques have become the conventional approach to base TTS models on, seeing how they can capture more complex structures in the input data. Not only has the deep learning approach helped create more natural-sounding speech, but it has also lowered the amount of necessary data and training time considerably. Some of the earliest and most known deep learning based TTS models are Tacotron [24] and WaveNet [25], developed by researchers at Google and DeepMind respectively. Tacotron is a sequence-to-sequence model, whereas DeepMind

is an autoregressive model. These models have paved the way for better models to be developed since their release in 2016-2017, and they are still used as baseline models to compare results with when conducting TTS research.

Speech synthesis belongs to a field that is progressing rapidly due to new focus, ongoing research, and advancements in deep learning techniques. Therefore, it is difficult to provide an accurate state-of-the-art (SOTA), as it changes fast because of new models being released. However, some of the most known new and improved deep learning models that have been released since the two previously mentioned are Tacotron2 [26], FastSpeech2 [27], FlowTTS [28], GlowTTS [29] and Transformer TTS [30]. An example of a newer model released in 2023 is Matcha-TTS [9], which was described in detail in Subsection 2.2.2.

3.2 TTS in Low-resource Languages

Making a TTS model for low-resource languages is a challenging task, as there might be very limited speech data available, and most models require a lot of data to achieve a good result. However, researchers are always looking for new approaches to deal with these challenges. This section will reference some of the relevant research for TTS development in low-resource languages and child speech.

Erica Cooper's 2019 thesis, "Text-to-Speech Synthesis Using Found Data for Low-Resource Languages", explores the potential of using existing recordings, also called "found" data, like radio broadcasts and audiobooks, to create TTS voices for languages that lack the extensive, high-quality datasets typically required for TTS development. Through comparative analysis, Cooper identifies the acoustic and prosodic characteristics that make found data viable for TTS purposes, despite its original intention for other uses. The research further investigates methods for selecting and adapting this "found" data to develop natural-sounding and intelligible TTS voices. By conducting a series of experiments that assess both the subjective and objective qualities of the synthesized voices, Cooper demonstrates that with careful selection and adaptation, found data can indeed be used to produce TTS voices of reasonable quality for low-resource languages [14].

In their 2023 report, Nouza et al. explain the development of a state-of-the-art end-to-end ASR system for Norwegian, addressing the language's complexity due to its many dialects and two written standards. To overcome the limitations of existing speech corpora, the team gathered extensive additional data from diverse sources, including broadcast and parliament

archives, YouTube, podcasts, and audiobooks. This effort resulted in a final model trained on 1,246 hours of Norwegian speech, further enhanced by transfer learning from a Swedish model. Although this article discusses an ASR system, it still deals with a lot of the same issues as a TTS system would when considering a low-resource language [31].

In their 2021 study, Byambadorj et al. propose innovative approaches for developing a TTS system for low-resource languages, focusing on Mongolian as a case study. They tackle the challenge of limited speech data by utilizing just 30 minutes of target language data. The researchers employed three strategies to train their TTS models: cross-lingual transfer learning, data augmentation, and a combination of both. Cross-lingual transfer learning involved leveraging high-resource language datasets in English and Japanese to improve the TTS model's performance in the target language. The combined approach integrated both methods, yielding the most natural-sounding synthesized speech according to subjective evaluations. Their findings demonstrated that using both cross-lingual transfer learning and data augmentation, particularly in a multi-speaker model format, resulted in superior TTS model performance [32].

3.3 Child Speech TTS

Similar to low-resource languages, child speech TTS models are challenging to develop due to a lack of available speech datasets. Additionally, child speech is very different from adult speech, as discussed in Subsection 2.5, which makes it difficult to synthesize. In Norwegian specifically, there is very limited previous research conducted regarding child speech TTS. However, there is previous related research in other languages, mostly English, that this thesis can fetch knowledge from.

In their 2022 article, Terblanche et al. provide a comprehensive analysis of the challenges and advancements in speech synthesis systems aimed at generating child voices. Acknowledging the traditional focus on adult speech, the scoping review covers studies from 2006 to 2021, highlighting the complexities in child speech synthesis due to the acoustic variability and articulatory errors unique to child speech. The review of 58 studies reveals efforts to adapt adult speech models using various techniques to produce child-like speech, which is notably more challenging than adult speech synthesis [33].

In their study from 2012, Begnum et al. explore creating high-quality synthetic Norwegian child voices by adapting a small amount of recorded child

speech data to an adult master voice. The initiative addresses the scarcity of commercial synthetic child voices due to the high production challenges and costs. The Norwegian child speech data was gathered in a professional studio, with an eleven-year-old boy as the target speaker. The project experimented with various strategies, including adjusting adult voice parameters and utilizing concatenative and formant synthesis techniques, but these either produced cartoon-like voices or were unsuitable for children's communicative needs. The focus shifted to statistical model-based technology, specifically HMM-based synthesis, allowing the adaptation of adult models to child-like voices using limited child speech data, significantly lowering the cost and complexity of production [34].

In their 2022 study, Jain et al. tackle the relatively underexplored domain of child speech synthesis within TTS research, which predominantly focuses on adult speech data. They introduce a novel training pipeline aimed at fine-tuning state-of-the-art neural TTS models with child speech datasets, leveraging a multi-speaker TTS retuning workflow for transfer learning. A significant contribution of their work is the cleaning and utilization of a publicly available child speech dataset to form a curated subset of approximately 19 hours, which serves as the foundation for their fine-tuning experiments [35].

In the 2023 study by Yiwere et al., the authors address the significant challenge of generating child-like speech data from adult speech, motivated by the scarcity of children's speech datasets for training speech-based artificial intelligence (AI) systems. They propose a novel speech augmentation pipeline that employs a phase vocoder-based toolbox for manipulating sound files, enabling the transformation of adult speech characteristics to resemble those of children. This involves adjusting the pitch and duration of adult speech utterances, aiming to make them sound more child-like [16].

Chapter 4

Methodology

This section will provide all the necessary information to conduct the same research as this thesis, such as detailed information about the testbed and the tools, the TTS architecture used to train the model, the data, how the models were trained, and how the model was evaluated. The section will only provide relevant information and statements motivating the choices, not discuss the consequences of the various choices, which will be saved for the discussion in Chapter 6.

4.1 Testbed & Tools

4.1.1 Computational Resources

Training a TTS model using only CPUs presents significant challenges due to their performance limitations and memory constraints, which can lead to frequent errors or excessively long training times. To address these issues, it is essential to utilize GPUs, which are specifically designed for computation-intensive tasks such as training TTS models. GPUs provide a robust architecture optimized for parallel processing, significantly enhancing computational efficiency. This improvement not only accelerates the training process but also reduces operational costs. By leveraging GPU capabilities, it becomes possible to conduct more extensive experiments within practical time frames.

SimulaMet, which is the research institution that this master thesis collaborates with, provides access to GPUs such as NVIDIA V100, A40, and A100 through their experimental eX3 cluster. Because it is a cluster meant to support numerous researchers, it employs Slurm as a resource manager. Slurm

organizes computational resources by creating a queue for jobs. This implies that you will need to schedule your job in the queue and wait, depending on the demand from other researchers. Consequently, training a model could require some patience since the resources you need might not always be immediately available

CPU	GPU	Memory	Architecture
64-core AMD EPYC 7763	NVIDIA A100 PCIe	40GB HBM2	x86_64 (AMD64)

Table 2: Relevant hardware of the GPU that trained the TTS model.

4.1.2 Data Curation

To create the datasets and train the models, several tools were also used. The tools were chosen based on necessity, simplicity, and quality, but other tools could be chosen instead.

Audacity, which is a free open-source tool to edit audio and record sounds, was frequently used when developing the datasets.[1] Specifically, it was used to split the data, change sample rate, and record speech. When recording speech, an Audio Technica AT2020USB+ microphone was used to capture a clear and high-quality voice.

Transcriptions of the speech data are necessary for the model to be able to map the speech to the corresponding text during training. To avoid manually transcribing the speech data, a Whisper model fine-tuned to Norwegian was used to automatically transcribe it [36].[2] Whisper, which is an acronym for WSPSR, short for Web-scale Supervised Pretraining for Speech Recognition, is an English ASR model developed by OpenAI [37]. It is commonly used as a base model to fine-tune into different languages, such as Norwegian.

Lastly, Microsoft Forms was used during the evaluation process as a tool to perform a qualitative assessment of the final model.

4.2 Matcha-TTS Model

The model architecture that was used to train the Norwegian TTS models is called Matcha-TTS[3] [9]. Matcha-TTS is a new deep learning based TTS

[1] https://www.audacityteam.org/
[2] https://huggingface.co/NbAiLab/nb-whisper-small-beta
[3] https://github.com/shivammehta25/Matcha-TTS

architecture introduced in 2023, and it was chosen for its fast and high-quality speech synthesis and its suitability in low-resource language scenarios. The models trained in this thesis were based on both single-speaker datasets and a multi-speaker dataset. Consequently, both the single-speaker and multi-speaker modes that Matcha-TTS provides were used during model training. Subsection 2.2.2 describes the architecture and design of Matcha-TTS in more detail.

4.3 Data

This section will describe the data used to train the models and how it was prepared.

4.3.1 Data Collection

As discussed in Section 2.3.2, collecting high-quality data for low-resource languages is a challenge. A publicly available Norwegian TTS dataset consisting of nearly 8 hours of speech by a single male speaker with Bokmål dialect used to exist.[4] Unfortunately, it was made publicly unavailable due to strict GDPR rules. The fact that there was only one speaker raised concerns regarding the anonymity of the speaker. Therefore, this thesis tests a few alternative approaches and compares the results to see which method works the best. The three approaches, which were chosen based on availability, simplicity, and potential, are ASR data, audiobook data, and self-recorded data.

4.3.1.1 Dataset 1 - Automatic Speech Recognition

There exist several high-quality Norwegian ASR datasets. Unfortunately, as discussed in Section 2.3.2, ASR datasets are usually not the best type of speech data for training TTS models. Nevertheless, due to the large amount of available ready-to-use data, it seemed like a good place to start. The ASR dataset that has been used is called NPSC[5], which is short for *Norwegian Parliamentary Speech Corpus* [38]. As the name suggests, it is a collection of speeches recorded during meetings with the Norwegian Parliament from 2017–2018. The entire dataset consists of about 140 hours of speech from

[4] https://www.nb.no/sbfil/talesyntese/NST-tts-dataset-documentation.pdf
[5] https://huggingface.co/datasets/NbAiLab/NPSC

267 unique speakers. That corresponds to 65000 utterances, which are all transcribed automatically, followed by manual proof-checking to correct mistakes in translation to ensure consistency and accuracy. However, 140 hours of speech is not necessary for the scope of this experiment. Therefore, only a small part of the dataset was used, corresponding to about 3 hours of speech data. These 3 hours include a total of 28 different speakers, making it a multi-speaker dataset. Figure 5 shows the audio duration distribution of all the clips and word count for the transcriptions. The average audio duration for the dataset is 6.68 seconds, and the average number of words per sentence is 18.02.

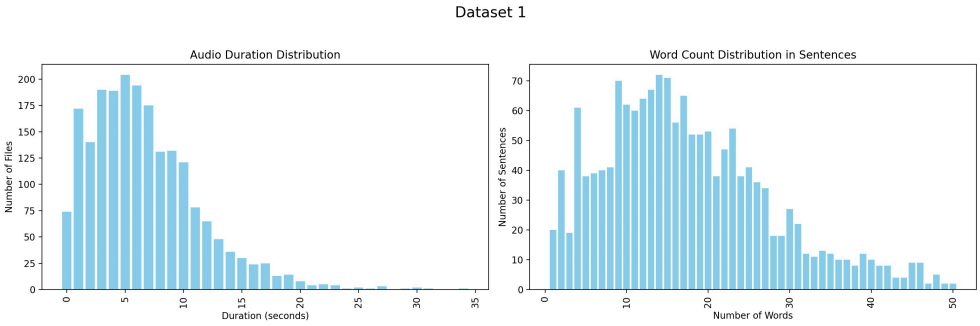

Figure 5: Audio duration distribution and word count distribution for all audio files and transcriptions for Dataset 1.

4.3.1.2 Dataset 2 - Audiobook

The audiobook data represents the ”found” data category used in this thesis. Found data is a term used for data that can be gathered by utilizing public sources, such as videos from YouTube or other streaming services, or sounds from radio broadcasts, audiobooks, or other sources. Even though this audiobook data was not originally intended to be used as training data for a TTS model, there is a general norm that everything publicly available can be used within certain constraints. Especially since this model will be used for academic and not commercial purposes, the ethical aspects have been considered. In total, about 1 hour of speech from a single male speaker with a Bokmål-dialect was gathered from an audiobook, which translates to a total of 713 utterances. Due to the lack of free single-speaker audiobooks with the desired amount of speech available, the chosen book is a low-level children’s audiobook. Therefore, the sentences in the datasets are relatively easy in terms of vocabulary, structure, and length, with most sentences being about 1–5 seconds long. That being said, the speech in the audiobook did not switch

between voices based on characters, which is a potential drawback of many audiobooks. As mentioned, the audiobook used in this thesis is free, but an alternative is to use one of the paid services to get access to audiobooks if it aligns with the budget. Figure 6 shows the distribution of the duration of audio clips and word count for the transcriptions. The average audio duration for the dataset is 1.94 seconds, and the average number of words per sentence is 9.57.

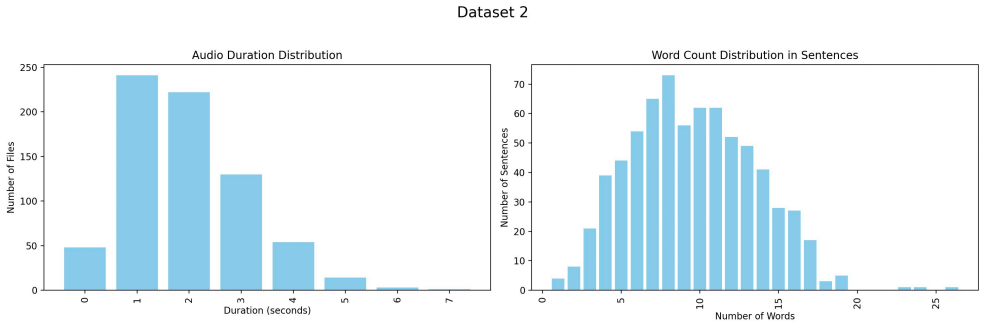

Figure 6: Audio duration distribution and word count distribution for all audio files and transcriptions for Dataset 2.

4.3.1.3 Dataset 3 - Self Recording

The final data collection method involved recording one's own voice. This approach ensures no ethical or legal issues due to the same reasons mentioned previously, and it eliminates the need for obtaining consent from others. Audacity was used for the recordings, utilizing its built-in silence finder tool to automatically split the audio into separate files at intervals of at least one second of silence. To ensure compliance with copyright regulations, all the recorded sentences were generated by ChatGPT. The prompts that instructed ChatGPT to generate text emphasized creating sentences with varying degrees of difficulty by altering sentence length, structure, vocabulary, and grammar. In total, almost 2 hours of speech were recorded, translating to a total of 989 utterances. Figure 7 shows the distribution of the duration of audio clips and word count for the transcriptions. The average audio duration for the dataset is 3.79 seconds, and the average number of words per sentence is 11.71.

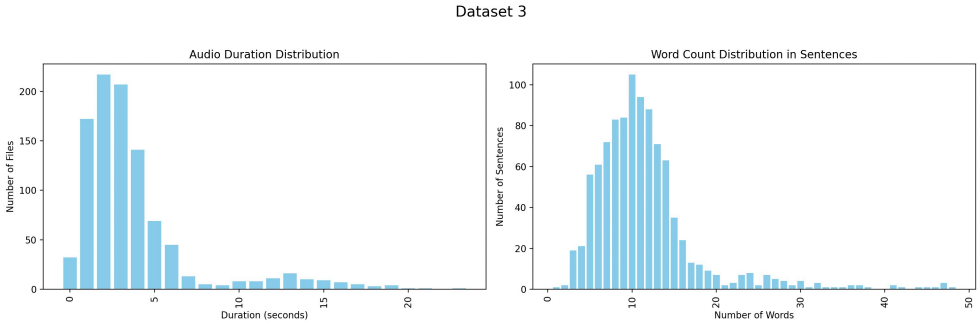

Figure 7: Audio duration distribution and word count distribution for all audio files and transcriptions for Dataset 3.

4.3.1.4 Data Quality

Mel-spectrograms are visual representations of the spectrum of frequencies in a sound signal over time. During the generation of the synthesized speech, Matcha-TTS also generated its Mel-spectrogram, and some examples can be seen in Figures 8, 9, and 10. The figures show the Mel-spectrograms for the same eight-second-long utterance, where the x-axis shows the time, the y-axis shows the pitches from low to high, and the colors show how loud the pitch is at each moment. The brighter the color, the louder the sound, meaning blue represents quiet sounds while yellow represents loud sounds. Among other things, they are useful for the analysis and identification of patterns in sounds and noise detection. Therefore, it is possible to detect the noise levels in the synthetic speech generated by the different models. Looking at how clear the patterns are and looking for scattered patches of color gives a good indication of how noisy the speech is, as clear lines or shapes usually indicate clean sounds with less noise, while random spots of color all over the place usually mean more noise. Figure 8 has scattered patches of color and a less clear pattern, resulting in a more messy Mel-spectrogram, indicating some background noise. Figure 9 has less scattered patches of color and a bit clearer pattern, indicating less background noise than the previous one. Lastly, Figure 10 has the least amount of scattered spots and a clearer pattern, indicating the smallest amount of background noise out of the three Mel-spectrograms.

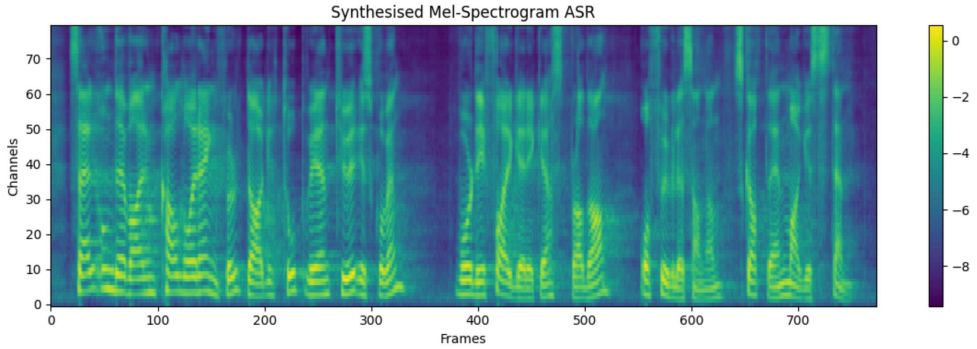

Figure 8: Mel-spectrogram for the synthesized speech generated by the model trained on ASR data.

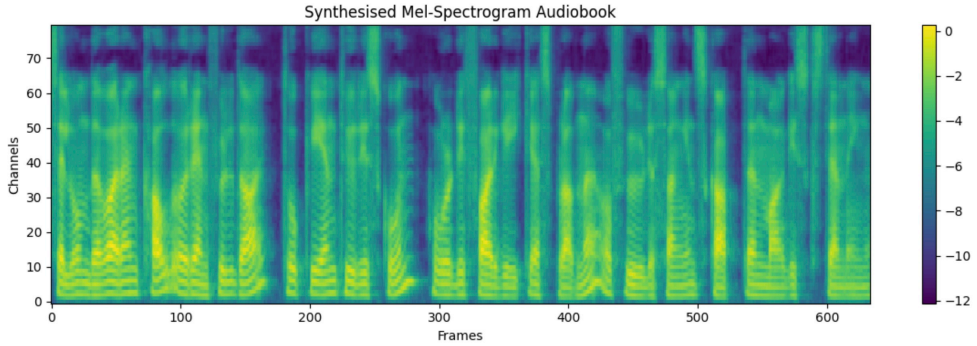

Figure 9: Mel-spectrogram for the synthesized speech generated by the model trained on audiobook data.

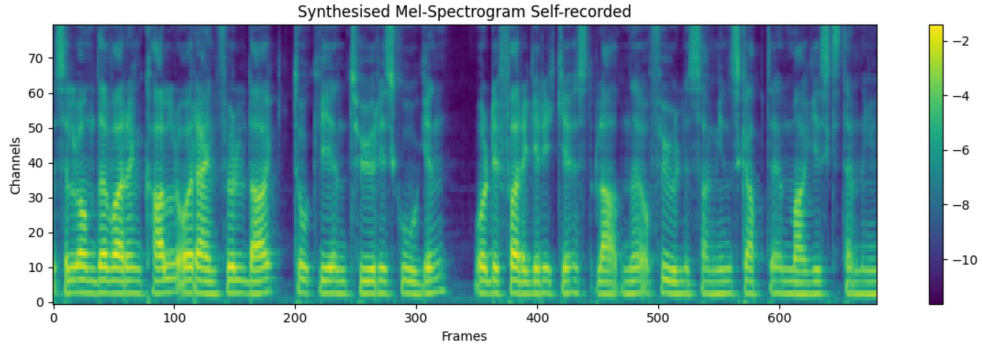

Figure 10: Mel-spectrogram for the synthesized speech generated by the model trained on self recorded data.

4.3.1.5 Dataset Overview

To consolidate the values of the datasets for easier comparison, all the most relevant statistics are summarized in Table 3. In addition to the values mentioned in the previous subsections, the table includes information about the Mel-spectrogram mean value and the Mel-spectrogram standard deviation value for each dataset. Among other things, Mel-spectrograms capture the intensity of the pitches and noise in audio signals, which are factors that can determine how suited the audio is for TTS development based on TTS dataset characteristics discussed in Section 2.3.1. Therefore, the Mel-spectrogram mean indicates the typical energy distribution in terms of average loudness across all audio files in the dataset. On the other hand, the Mel-spectrogram standard deviation indicates the variance in the Mel-spectrogram values across the datasets. A higher value indicates variability in terms of energy levels in the audio signal, while a lower value indicates more consistency.

Name	Dataset 1	Dataset 2	Dataset 3
Data Type	ASR	Audiobook	Self Recorded
Number of Speakers	28	1	1
Total Utterances	1588	713	989
Mean Audio Duration (Seconds)	6.68	1.94	3.79
Mean Sentence Word Count	18.02	9.57	11.71
Mel Spectogram Mean	≈ -5.132	≈ -5.681	≈ -7.267
Mel Spectogram Standard Deviation	≈ 1.636	≈ 2.487	≈ 1.878

Table 3: Relevant information about the three datasets

4.3.2 Pre-processing

After collecting the data, the next step is the pre-processing phase, which involves transforming the data into a dataset suitable for training a TTS model. The pre-processing also standardizes the datasets as much as possible. Depending on the data, various measures must be taken to prepare the dataset, such as changing the sample rate, audio splitting, and formatting. Each subsection will explain a pre-processing step that was applied to at least one

of the datasets. Several tools were used in the various pre-processing steps, which can be read more about in Subsection 4.1.

4.3.2.1 Audio Splitting

To begin with, one large audio file was transformed into multiple smaller files by splitting it into individual files for each utterance using Audacity. The built-in silence finder tool was used to identify pauses, which represented a new sentence. By using this automatic tool, the manual effort of splitting the sentences was avoided. This process was necessary only for Dataset 2 and Dataset 3, as Dataset 1 was already divided into separate files. As mentioned in Subsection 2.3.1, a high-quality dataset should have a normalized distribution of audio clip lengths and corresponding transcriptions to ensure coverage of both short and long sentences. Since most sentences in daily speech are of medium length, the dataset should predominantly cover medium-length sentences. However, as the audio was split based on utterances rather than length, this factor had to be considered during the recording process.

4.3.2.2 Sample Rate

The sample rate of audio is the number of samples that are taken from a continuous signal per second to create a digital signal, and it is measured in Hertz (Hz). Applications use various values of sample rates based on available resources and purpose. Audio used for TTS training usually has a sample rate between 22050Hz and 48000Hz, in contrast to ASR systems which usually use 16000Hz. That is because TTS systems need to be able to capture detailed acoustic characteristics to accurately generate synthetic speech from text.[6] However, a sample rate of more than 22050Hz is often only preferred in situations where the goal is to create a professional application and there is access to professional equipment and studio. In addition, using 22050Hz reduces the size of the audio files, which in turn decreases the storage and computational requirements for TTS training. Therefore, all the audio used in the experiments has a sample rate of 22050Hz.

There are various methods to change the sample rate, either through coding a script or using other available tools. For Dataset 2 and Dataset 3, Audacity was used not only to split the audio files but also to change the sample rate to 22050Hz. In contrast, Dataset 1, which was already prepared and split into audio files, had its sample rate changed using a small Python script.

[6] https://www.futurebeeai.com/blog/sample-rate-for-asr

4.3.2.3 Formatting

An essential step in preparing the dataset for training is formatting it in a way that the TTS model can understand. The data consists of thousands of audio files generated during the audio splitting step and stored in a data folder. However, for the model to learn and recognize the text associated with specific sounds for future synthetic speech generation, it needs access to both the audio files and their corresponding transcriptions. Therefore, TTS models utilizes file lists, which are text files that map audio files to their transcriptions. Each dataset has its own file list, formatted as follows:

```
path/to/audio|This format is for single-speaker datasets.
path/to/audio|id|This format is for multi-speaker datasets.
```

As illustrated, each line in the file list corresponds to a single audio file. In a single-speaker dataset, as shown in the first line, the file list maps audio to text by placing the path to the audio file and the corresponding transcription on the same line, separated by the delimiter "|". For a multi-speaker dataset, as demonstrated in the second line, an additional number representing the speaker's ID is included between the audio path and the text, also separated by the "|" delimiter. This inclusion is crucial, as the model needs to distinguish between different voices for accurate reproduction during the generation of synthetic speech in later stages.

A Python script was developed to create the different file lists. This script iterates through the data folder containing all the audio files for each specific dataset and employs a Whisper model fine-tuned to Norwegian to automatically transcribe the audio files. Although the Whisper model has a WER of 8.3%, indicating some risk for transcription errors, it significantly reduces the time and effort required for manual transcription. Once an audio file is transcribed, the script appends the transcription to the file list as a new line. Only Dataset 2 and Dataset 3 required transcription using Whisper, as Dataset 1 already had available transcriptions. An alternative approach was tested on Dataset 2: the original text from the audiobook was pasted into a single file, and a Python script was used to iterate through all the speech samples, pairing them with the text sequentially. While this method could theoretically work, the audio splitting for the audiobook data was imperfect due to the reader's varying pauses, leading to concatenated speech samples and disrupted pairings of speech and text sentences. Another potential method to ensure more accurate transcriptions would involve using Whisper to transcribe initially, then writing a script to iterate through all transcriptions, calculating the similarity between the transcribed sentences and the original sentences.

If a transcribed sentence did not match any original sentences, it could be replaced with the most similar sentence. However, the effort required for this approach was deemed unworthy considering the potential benefits.

Each dataset is divided into two sets: a training set and a validation set. The training set is utilized for training purposes, whereas the validation set is employed throughout model training to monitor progress and assess how effectively the dataset is responding to the training. Consequently, file lists are generated for both the training and validation sets during the formatting process for each dataset.

4.3.2.4 Mactha-TTS Modifications

Training a TTS dataset with the Matcha-TTS architecture using a self-made dataset in a new language required several code adjustments. The most notable changes are highlighted here.

TTS architectures use cleaners to pre-process input text before it is fed into the model, aiming to normalize, simplify, and ensure consistency across the data. Introducing a new written language brings new characters and symbols that are not included in the default vocabulary of these cleaners. To address this, a script was implemented to iterate through all sentences in the dataset and add the missing characters.

Additionally, training both single-speaker and multi-speaker models presented architectural challenges related to embeddings and layers. The embedding dimensions had to be modified to prevent embedding mismatch errors and ensure layer compatibility.

4.4 Model Training

A total of four models were iteratively trained using the three datasets: one model for each dataset and one model using all datasets combined. Both the single-speaker and multi-speaker modes were used during model training as Dataset 2 and Dataset 3 were single-speaker, while Dataset 1 was multi-speaker. The objective was to develop the best possible Norwegian TTS model by experimenting with different datasets. This iterative approach allowed for comparative analysis of the results, aiding in the identification of the most effective data collection method for a low-resource language. This was particularly crucial given the absence of open-source Norwegian TTS models capable of producing comparable synthetic speech. After each model completed its training, decisions were made on the next steps to enhance

the TTS model's quality, including selecting the most appropriate dataset for further training. In other words, the different datasets were developed consecutively based on the previous model's results.

All models were trained using the default hyperparameters of MatchaTTS, as provided in the GitHub repository. These default parameters are designed to perform well across various datasets. The decision to keep the hyperparameters consistent across all model training iterations was driven by the primary focus on evaluating different datasets to identify the most suitable one for TTS development. Changing both hyperparameters and datasets simultaneously would make it more difficult to determine which factors had the most significant impact on the results.

The models were trained for various amounts of time, as this was one of the factors that were adjusted between the training iterations to improve the models.

Matcha-TTS uses Mel-spectrograms during training as a part of its feature prediction process. The model predicts durations, which are used for upsampling the encoder's output vectors to obtain the predicted average acoustic features, including Mel-spectrograms. These features condition the decoder, guiding the synthesis process while ensuring efficient and high-quality speech generation. Unlike some other models, Mel-spectrograms in Matcha-TTS are not used as the mean for initial noise samples, simplifying the training and synthesis steps.

4.4.1 Model 1 - Automatic Speech Recognition

Model 1 was trained on Dataset 1 for 2 days. This dataset was chosen for the first experiment because it already included prepared audio and transcriptions. Moreover, since Dataset 1 was originally developed for ASR training, it was not anticipated to perform well in TTS tasks, as discussed in Subsection 2.3.2, making it an ideal starting point. Since Dataset 1 was a multi-speaker dataset, it was trained in the multi-speaker mode that the Matcha-TTS architecture provides. To achieve that, the data had to be prepared using a file list designed for multi-speaker training, as described in section 4.3.2.3. Additionally, code had to be adjusted in the file describing the data, such as specifying the amount of speakers in the dataset, enabling Matcha-TTS to train in multi-speaker mode by applying more embeddings and layers.

4.4.2 Model 2 - Audiobook

Model 2 was trained on Dataset 2 for 3 days. Following the results from Model 1, it was decided to create a new dataset with a focus on data more suitable for TTS, despite the need for additional preprocessing. Audiobook speech was chosen as the source for this dataset due to its availability and quality. Once Dataset 2 was developed, the code was adjusted to handle a single-speaker dataset, and the training for Model 2 commenced.

4.4.3 Model 3 - Self Recording

Model 3 was trained on Dataset 3 for 1 week. Building on the results from Model 2, creating another dataset became a priority. Given the thesis' focus on developing a PoC TTS model for low-resource languages, the accessible approach of recording one's own voice was chosen. This method mirrors audiobook speech, as both feature a single speaker and controlled, planned speech. It also allows for greater control over variables such as sentence length and the quantity of data.

4.4.4 Model 4 - Combined Dataset

Model 4 was trained on a combination of all three datasets for 3 days. Due to time constraints, the training time had to be lowered compared to Model 3. The objective was to assess the performance of a TTS model trained on a more extensive and diverse dataset compared to individual datasets. Since the goal is to develop a PoC TTS Model with as good intelligibility and naturalness as possible, it was also decided to perform additional pre-processing on Dataset 1 before training Model 4. As background noise was one of Dataset 1's drawbacks, we developed a Python script to reduce noise using the *noisereduce* library. The expectation was that combining the datasets would improve the model's ability to learn more patterns, thereby enhancing naturalness and intelligibility. This required creating an additional file list to map the combined datasets, although the overall file structure remained unchanged. Since this model was trained on a combination of all datasets, it followed the same steps as Model 1 in terms of preparing it for multi-speaker training.

4.5 Evaluation Method

This section describes the evaluation method, explaining how the evaluation was performed and how the scores were calculated. Consequently, the results will be presented and analyzed separately in Chapter 5. Evaluating a model's performance is crucial in AI model development, as it demonstrates the effectiveness of the applied methodology. Future research can either adopt or avoid the methods used in this thesis based on the evaluation outcomes. However, as discussed in Section 2.4, evaluating a TTS model presents several challenges. The models will be assessed using a mixed-methods approach, combining both quantitative and qualitative evaluations. All models will have their intelligibility objectively evaluated using an ASR model to calculate the WER of synthesized speech samples. Additionally, Model 3 and Model 4 will undergo subjective evaluations of intelligibility and naturalness by calculating MOS through two iterations of user surveys. This combination of objective and subjective evaluations allows for a comprehensive comparison of results, facilitating conclusions on the most effective methodologies. All the evaluations required the generation of synthetic speech. Matcha-TTS allows different parameters when generating speech, such as temperature, speaking rate, steps, and denoiser strength. Temperature refers to how much randomness should be applied when generating the speech, and steps refer to the number of iterations the generation takes. However, due to consistency in the speech generation, the default parameters were applied.

4.5.1 Objective Evaluation Method

To objectively evaluate the intelligibility of the models using an ASR model to calculate the WER, 100 sentences were generated by ChatGPT and saved in a text file, with each sentence on a separate line. Figure 11 displays the audio duration distribution and word count distribution for these sentences. The sentences were crafted to vary in length, structure, vocabulary, and context, ensuring broad phonetic coverage and incorporating various prosodic features. This approach aimed to ensure reliability and minimize bias towards any specific model. Only sentences ending with a period were used, maintaining consistency with the focus of the data used to train the models. The objective evaluation was conducted over three iterations.

Iteration 1 focused on testing the various speakers of Model 1 to identify the best-performing speaker, as it is a multi-speaker model. To achieve this, the five speakers with the most utterances in the dataset synthesized 100

speech samples each from the sentences generated by ChatGPT. The Whisper ASR model, previously used for data formatting in Section 4.3.2.3, then transcribed all the generated speech samples, creating a new text file in the same format as the original. A Python script utilizing the JiWER library, which calculates metrics for ASR model evaluation, was used to determine the WER by comparing the original text file to the transcriptions. The speaker with the lowest WER was selected to represent Model 1 in Iteration 2, where the same process was performed for Model 2 and Model 3.

Iteration 2 replicated the test conducted in Iteration 1, but instead of comparing different speakers within the same model, it evaluated Model 1, Model 2, and Model 3 against each other. Since Model 1 had already undergone the test and its WER was calculated, only Model 2 and Model 3 were subjected to the evaluation process.

Iteration 3 also followed the same procedure as Iteration 1, but this time it evaluated Model 4. Since Model 4 is a multi-speaker model trained on a combined dataset, the speakers were selected to match those evaluated in Iteration 2. This approach aimed to determine if there was any improvement in intelligibility after training on the combined dataset.

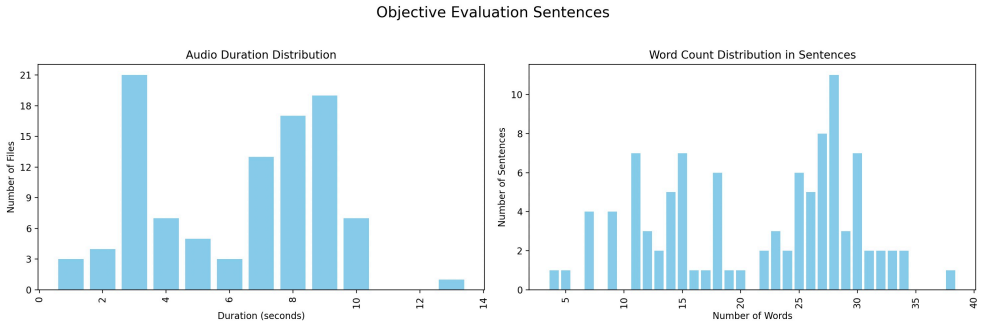

Figure 11: Audio duration distribution and word count distribution for the 100 audio files and transcriptions for the sentences used in the objective ASR evaluation.

4.5.2 Subjective Evaluation Method

MOS is a subjective metric derived from user surveys, where participants provide their opinions on various speech samples. The average score from their responses indicates the general perception of the listeners. Conducting a MOS evaluation can be time-consuming for participants since they need to listen to many speech samples. To mitigate this, the user surveys were conducted in two iterations, focusing only on Model 3 and Model 4. Additionally, all information and questions were provided in Norwegian, the

target language of the TTS model, to make the evaluation process easier and more accessible for participants. As highlighted in Section 2.4, certain pitfalls must be avoided when conducting user surveys for MOS calculations, such as failing to provide sufficient details to participants and not adequately documenting the process, which can undermine the evaluation's validity. To address these issues, the user surveys were conducted according to the ITU standard for MOS interpretation and reporting, with a strong emphasis on clearly communicating all steps taken [39].

The same user survey was conducted for both Model 3 and Model 4, with the only difference being the speech samples generated by the respective models. The survey for Model 3 was conducted first, followed by Model 4. Microsoft Forms was selected to conduct the user surveys over Google Forms due to its superior integration with YouTube videos, which were used as a medium for presenting speech samples. The form was divided into two sections. The first section collected personal information to identify potential trends and made specific requests to ensure the consistency and validity of answers. The second section was dedicated to the actual evaluation of the speech samples. An introductory segment preceded the first section, providing participants with context. It explained what TTS is, the purpose of the evaluation, what the process would involve, and outlined requests for participants to consider during the evaluation. It also included an estimated completion time and assured participants that their responses would remain anonymous. The introduction aimed to motivate the participants to take the test thoroughly while feeling safe and enhancing their experience through a better understanding. The evaluation was estimated to take approximately 10 minutes, based on trials conducted by two independent subjects prior to the survey's release. The responses from these two subjects were excluded from the final evaluation.

Section 1 of the user study form aimed to identify trends and ensure the validity and consistency of responses by asking personal questions. To uncover trends, participants were asked about their age and frequency of using TTS technology, such as virtual assistants (e.g., Siri, Alexa) or GPS. Additionally, questions regarding hearing impairments, whether Norwegian is their first language, whether they would be using headphones, and whether they would be surrounded by noise while listening to the audio were included to maintain consistency and validity.

Section 2 began by explaining the evaluation process to the participants and explaining what they would be rating immediately after listening to the audio clips. Participants were presented with 15 synthetic speech samples

in the form of YouTube videos with a black background. The sentences used to generate these samples were produced by ChatGPT, with a focus on varying length, structure, vocabulary, and context. The prompt that instructed ChatGPT to generate the samples was: "*I need to evaluate my TTS model. Therefore, I need to create 15 sentences in Norwegian with varying lengths, vocabulary, structure, context, intonation, and other elements to make phonetically balanced sentences. I only want sentences that end with a period. Can you create 15 sentences of different lengths: 3 sentences with about 50 words, 3 with about 40 words, 3 with about 30 words, 3 with about 20 words, and 3 with about 10 words?*" This approach ensured broad phonetic coverage and included various prosodic features, aiming to replicate the phonetically balanced Harvard sentences in Norwegian.[7] Figure 12 illustrates the distribution of audio duration and word count for the speech samples used in the evaluation, with durations ranging from 2 to 16 seconds and word counts ranging from 9 to 53.

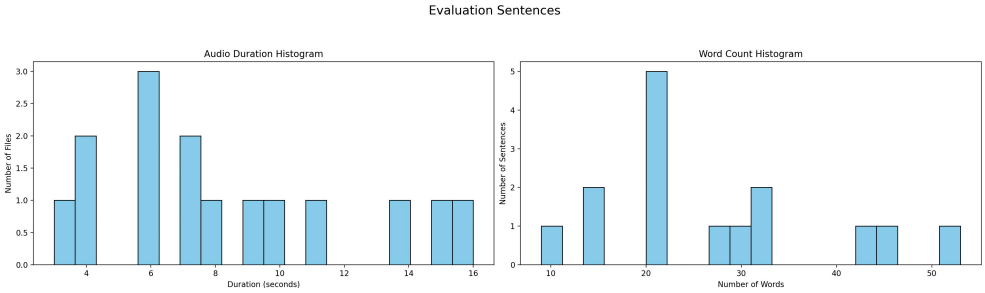

Figure 12: Audio duration distribution and word count distribution for the 15 audio files and transcriptions for the sentences used in the subjective MOS evaluation.

For each audio clip, the participant was asked to rate the intelligibility and naturalness of the speech on a Likert scale from 1-5. When rating the intelligibility, the participant was instructed to focus on how clear and understandable the speech was, with an emphasis on how easy it was to understand the spoken words. Similarly, when rating the naturalness, they were instructed to evaluate how much the synthetic speech resembles human speech, focusing on pronunciation, rhythm, and intonation. Specifically telling the participants what to focus on when rating the two factors improves the consistency across the answers. The following Likert scale from 1-5 was used, where 1 is the worst and 5 is the best, except for the words being

[7] https://www.cs.cmu.edu/afs/cs.cmu.edu/project/fgdata/OldFiles/Recorder.app/utterances/Type1/harvsents.txt

translated into Norwegian. The scale is in line with the guidelines in the ITU standard [39].

1. Bad
2. Poor
3. Fair
4. Good
5. Excellent

Lastly, to ensure the validity of the evaluations, three tests were randomly applied throughout the evaluation. The tests asked the participants to write the sentence they had just listened to, forcing them to pay attention to the audio and not rush through.

Chapter 5

Results and Analysis

The results chapter will provide the results of the objective and subjective evaluations described in Subsection 4.5. Further interpretation and discussion will be saved for Chapter 6.

5.1 Objective Evaluation

As described in Subsection 4.5.1, the objective evaluation only evaluates the intelligibility of the various TTS models. Intelligibility is evaluated using a WER score, which represents the percentage of wrong words in the transcription of the synthetically generated speech samples compared to the original text. The results in this section belong to the iterations described in Subsection 4.5.1.

5.1.1 Iteration 1 - WER Evaluation on Dataset 1

Model 1 was trained on Dataset 1, a multi-speaker dataset featuring 28 different speakers. To compare the WER scores among the three different models, a single speaker from Dataset 1 needed to be selected to represent Model 1. It is reasonable to assume that the speaker with the most utterances in the dataset would demonstrate the best intelligibility performance. However, to ensure accuracy, a WER comparison was conducted among the five speakers with the highest number of utterances in the dataset. Figure 13 presents the results, with the x-axis showing the speaker IDs as they appear in Dataset 1. The speakers are ordered from lowest to highest based on their number of utterances: Speaker 168 with 67 utterances, Speaker 32 with 106 utterances, Speaker 101 with 136 utterances, Speaker 190 with 205

utterances, and Speaker 207 with 325 utterances. These statistics, along with the corresponding WER scores, are summarized in Table 4. The substantial variation in the number of utterances among these speakers provides a solid basis for evaluating whether the quantity of utterances significantly impacts the intelligibility performance of the speaker.

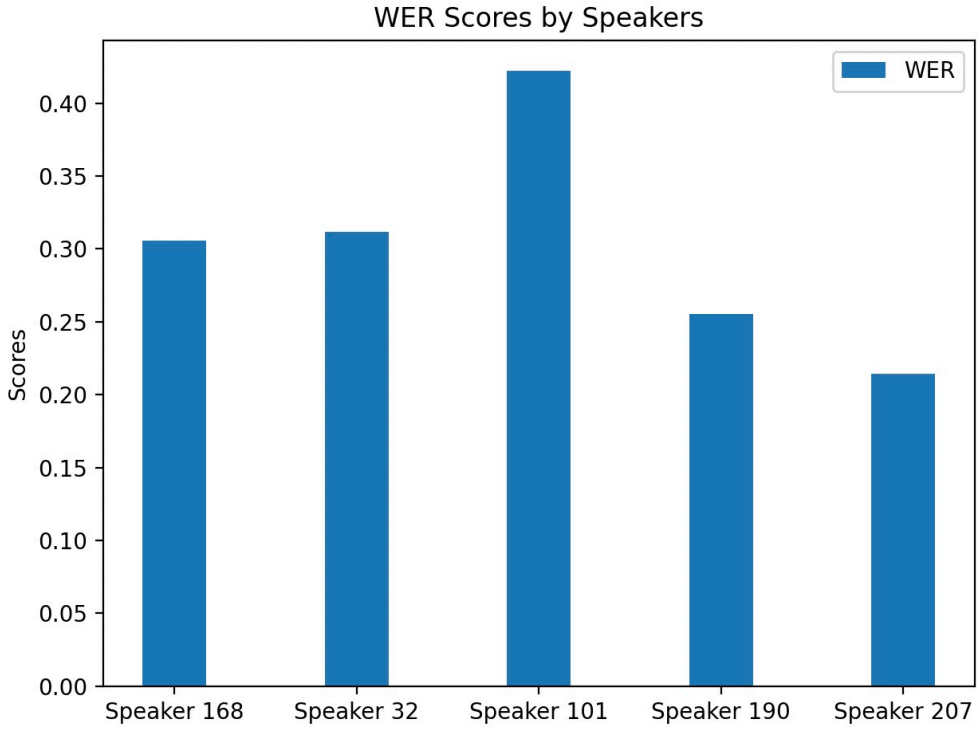

Figure 13: WER scores by speakers in Dataset 1.

	Speaker 168	Speaker 32	Speaker 101	Speaker 190	Speaker 207
Utterances	67	106	136	205	325
WER	30.56%	31.17%	42.22%	25.53%	21.44%

Table 4: A summary of the relevant values for the WER test, including the total amount of utterances, WER score, total errors, and the amount of substitutions, deletions, and insertions.

As illustrated in Figure 13, the higher WER scores for Speaker 101 and Speaker 32 compared to Speaker 168 suggest that the WER score does not

solely depend on the number of utterances in the dataset. Table 4 shows that both Speaker 101 and Speaker 32 have significantly more utterances than Speaker 168. This indicates that the quality of Speaker 168's utterances must be higher than those of Speaker 101 and Speaker 32.

To investigate further, the word count distribution among all the sentences for each speaker was analyzed, as shown in Figure 14. The x-axis represents word count in bins of 5-value intervals, while the y-axis represents the number of sentences. Outliers with a word count above 50 were removed to ensure consistency in the x-axis scale across all plots. Speakers 168, 190, and 207, who achieved the top three WER scores, exhibit a more normalized distribution of word counts, providing good coverage of short, medium, and long sentences. In contrast, Speakers 32 and particularly 101 show a distribution heavily skewed towards shorter sentences, lacking coverage for medium and long sentences. These comparisons lead to the conclusion that the word count distribution is as crucial, if not more so, to the intelligibility performance of the speakers in Dataset 1 as the number of utterances.

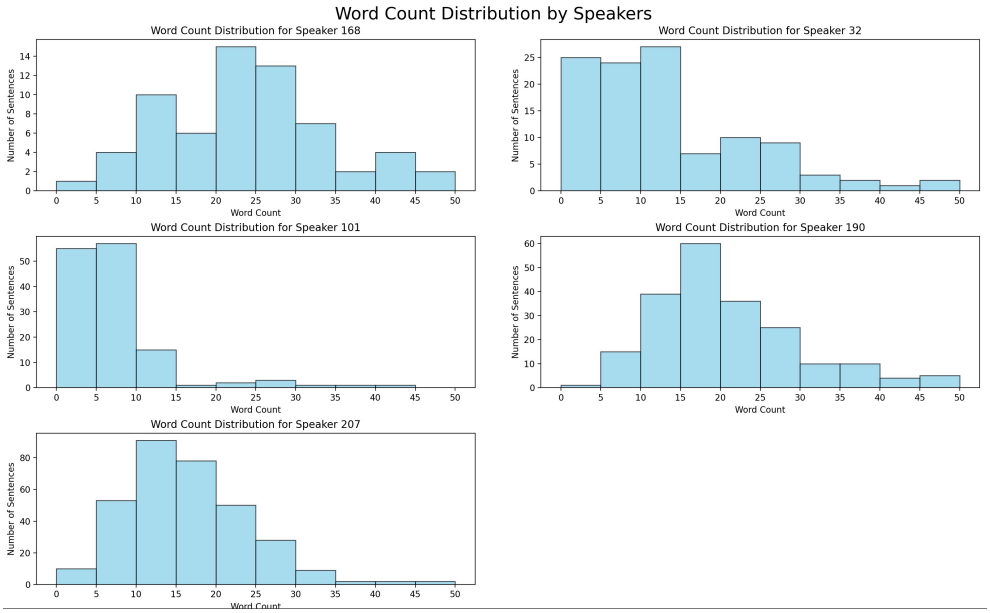

Figure 14: The word count distribution of all the sentences for each speaker in Dataset 1.

5.1.2 Iteration 2 - WER Evaluation on Model 1, 2 and 3

After identifying the most intelligible speaker in Dataset 1, a similar test was conducted to evaluate and compare the intelligibility of Model 1, Model 2, and Model 3. Since Speaker 207 achieved the lowest WER in Iteration

1, this speaker was chosen to represent Model 1 in Iteration 2. Figure 15 illustrates the WER scores for the three models, showing fluctuations in their intelligibility. Model 1 achieved a WER score of 21.44%, making it the second most intelligible model. Model 2, however, had a significantly higher WER of 31.92%, positioning it as the least intelligible model. In contrast, Model 3 demonstrated the highest intelligibility with a WER score of 17.25%. These WER scores, along with the distribution of errors, which will be described in the next paragraph, are numerically summarized in Table 5.

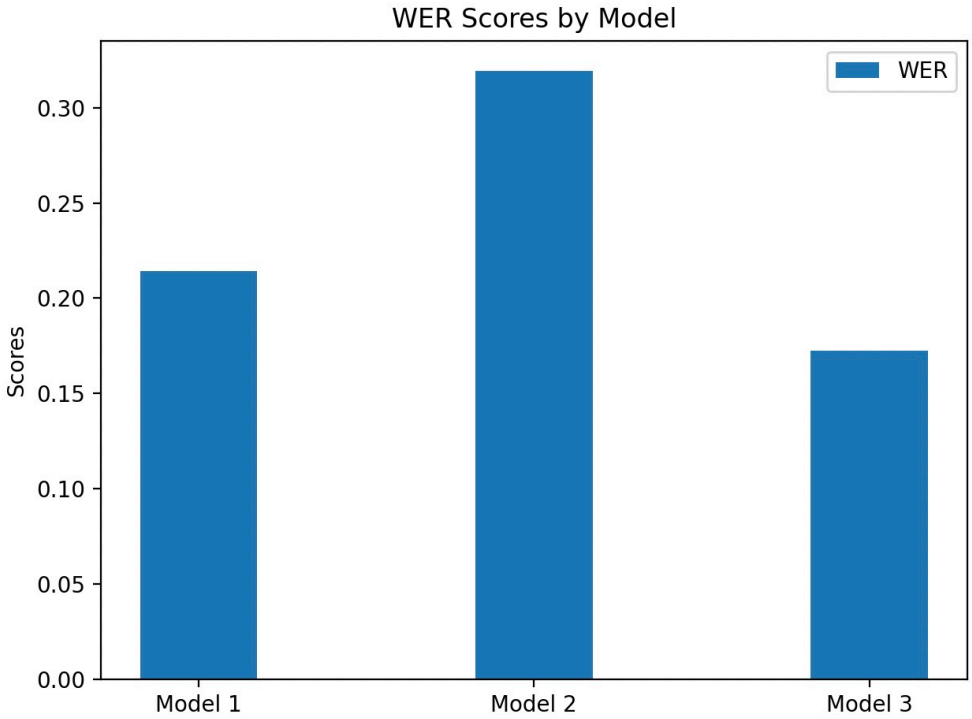

Figure 15: WER scores by model in Iteration 2.

Another valuable statistic to examine is the distribution of error types among the total errors for the different models. An error, which negatively impacts the WER score, can be one of three types: substitution, deletion, or insertion. A substitution error occurs when a word is transcribed incorrectly by the ASR model in place of the correct one. A deletion error indicates that a word is missing and therefore has not been transcribed. An insertion error means that an extra word has been transcribed by mistake. Figure 16 illustrates the distribution of these error types across the different models, while Table 5 provides a numerical summary. The 100 sentences used to generate the speech samples for WER calculation consist of a total of 2127 words. Model 1 had

a total of 456 errors, which included 329 substitutions, 109 deletions, and 18 insertions. Model 2 had a significantly higher total of 679 errors, comprising 272 substitutions, 380 deletions, and 27 insertions. Model 3 had the fewest errors, with a total of 367, consisting of 220 substitutions, 121 deletions, and 26 insertions.

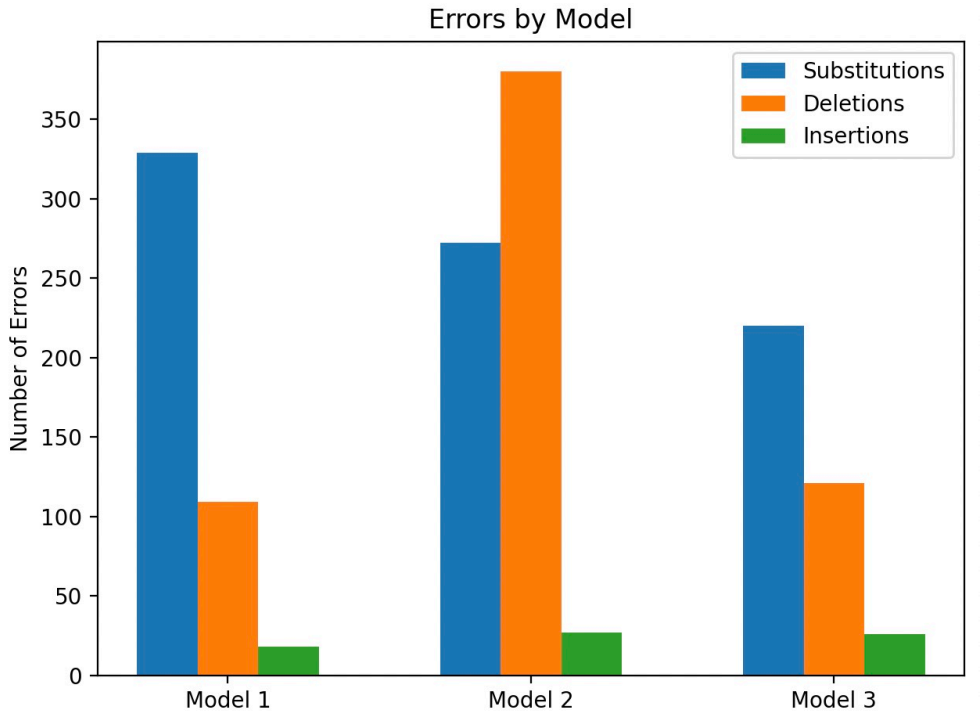

Figure 16: The distribution of errors by model among the categories substitutions, deletions, and insertions.

As illustrated in Figure 16 and detailed in Table 5, the distribution of errors varied significantly between the models. The number of insertions was relatively low across all models, indicating that the models rarely recorded additional words that were not present. However, Model 1 and Model 3 exhibited a much higher number of substitutions compared to deletions, while Model 2 had a notably higher number of deletions compared to substitutions. In practice, this means that Model 1 and Model 3 frequently replaced one word with another incorrect word, whereas Model 2 tended to omit the word entirely. This suggests that Model 1 and Model 3 had an easier time capturing the flow and natural pauses of speech but struggled with accurately identifying individual words. Conversely, Model 2 was better at recognizing individual words but had difficulty with the overall flow and natural pauses, often merging

multiple words into one.

	Model 1	Model 2	Model 3
WER	21.44%	31.92%	17.25%
Total errors	456	679	367
Substitutions	329	272	220
Deletions	109	380	121
Insertions	18	27	26

Table 5: The WER scores and distribution of errors by models among the categories substitutions, deletions, and insertions

5.1.3 Iteration 3 - WER Evaluation on Model 4

After evaluating Models 1, 2, and 3, Model 4 was trained using a combined dataset that included all three datasets. This approach aimed to determine if training on a larger, more diverse dataset would improve the intelligibility and naturalness of the synthesized speech. To enable direct comparison, the same tests conducted in Iteration 2 were repeated for Model 4. The speakers used to generate the speech samples were the same as in previous tests: Speaker 207 for Model 1, Speaker 1000 for Model 2, and Speaker 1001 for Model 3.

The results, depicted in Figure 17 and summarized in Table 6, show significant variation from the corresponding plots from Iteration 2. Speaker 207 achieved a WER of 27.41%, Speaker 1000 achieved a WER of 21.67%, and Speaker 1001 achieved a WER of 14.95%. This indicates a moderate downgrade of 5.97% for Speaker 207, a significant improvement of 10.25% for Speaker 1000, and a slight improvement of 2.3% for Speaker 1001. Consequently, the ranking of the models based on intelligibility changed compared to Iteration 2, with Speaker 207 and Speaker 1000 switching places.

These results suggest that Speaker 207, representing the multi-speaker ASR dataset, did not benefit from the additional data. While theoretically, more data should have been advantageous, the attempt to improve Dataset 1's quality in Model 4 by removing noise might have backfired and caused harder conditions for the model to learn the natural flow of speech, resulting in a downgrade. The denoising applied to Dataset 1, which is described in Subsection 4.4.4, was quite strong, which might also have lowered the

loudness of the speech on certain occasions. On the other hand, Speaker 1000, representing the audiobook dataset, significantly benefited from the combined dataset, utilizing the additional data to better learn basic speech patterns such as rhythm and flow. Lastly, Speaker 1001, representing the self-recorded dataset, also showed improvement, becoming more stable and learning additional patterns from the combined training data.

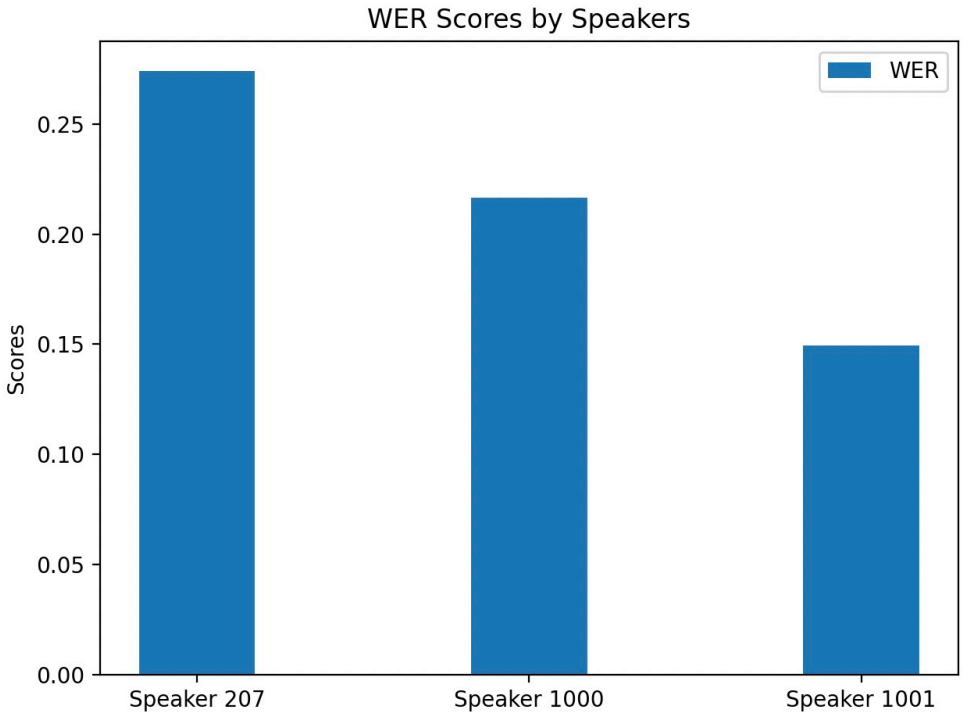

Figure 17: WER scores by speakers in Model 4, which is a model trained on a dataset combined of all three datasets.

The error distribution across the categories of substitution, deletion, and insertion is illustrated in Figure 18. The overall pattern is quite similar to the distribution of Iteration 2, with one notable exception: Speaker 1000. This speaker has significantly reduced the number of deletions, nearly equalizing the count of substitutions and deletions. This finding supports the earlier observation that Speaker 1000 has improved in learning fundamental speech patterns such as rhythm and flow. The reduction in deletions suggests that Speaker 1000 has become better at interpreting pauses and maintaining the natural flow of speech without omitting words.

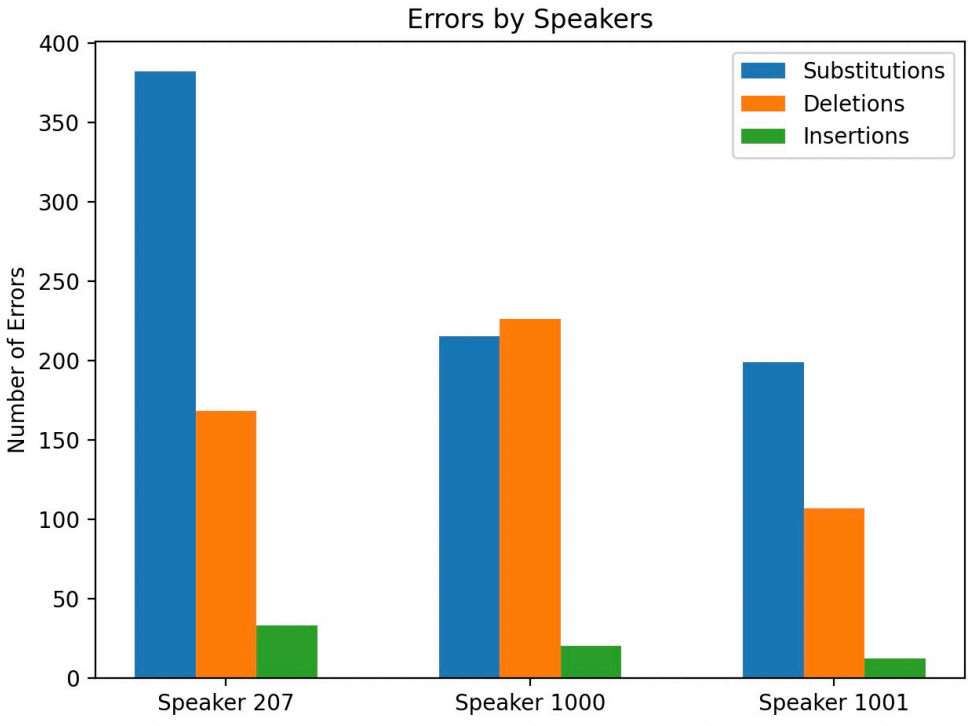

Figure 18: The distribution of errors by speakers among the categories substitutions, deletions, and insertions.

	Speaker 207	Speaker 1000	Speaker 1001
WER	27.41%	21.67%	14.95%
Total errors	583	461	318
Substitutions	382	215	199
Deletions	168	226	107
Insertions	33	20	12

Table 6: The distribution of errors by speakers in Model 4 among the categories substitutions, deletions, and insertions

5.2 Subjective Evaluation

The subjective evaluations were conducted in two iterations as two separate user surveys. The user surveys were identical, except for the speech samples,

which were generated by Model 3 in User Survey 1 and Model 4 in User Survey 2. Even though the speech was generated from different models, they still featured the same speaker, enabling a comprehensive comparison. This section will present the results from the user surveys and analyze them.

MOS was used to evaluate intelligibility and naturalness. Analyzing the results can uncover trends between ratings and factors such as how well the model performed with speech of different lengths. Since there were so many questions in the user surveys and therefore equally as many plots generated by Microsoft Forms, all the plots showing the results can be found for User Survey 1 and User Survey 2 in Appendix A and B, respectively. Therefore, the plots will be referenced but not shown in the upcoming sections.

5.2.1 User Survey 1: Evaluating Model 3

As shown in Figure 25, a total of 25 participants completed the user survey. The majority, 17 participants, belonged to the age group 24–30, while 4 participants were in each of the age groups 31–40 and 50+. Consequently, most responses came from individuals in their mid- to late twenties. Furthermore, Figures 26, 27, 28, 29, and 30 respectively illustrate the distribution of participants in terms of having a hearing impairment, having Norwegian as their first language, using headsets while listening to the speech, being in a noisy environment during the test, and the type of device used. The data indicates that only one participant did not have Norwegian as their first language. Additionally, no participants reported having a hearing impairment, which is crucial for the reliability of the responses. Moreover, 18 participants used headsets while 7 did not, and 22 were in a quiet environment while 3 were surrounded by noise. Finally, 3 participants used a PC, 11 used a Mac, and 11 used a phone for the survey. The fact that most participants were in a quiet environment and using headsets supports the consistency of the responses, although the diversity in device type might affect the results. However, given the limited number of responses, all answers were accepted as valid.

With 25 participants rating 15 speech samples each, this resulted in 375 evaluations regarding the quality of both intelligibility and naturalness of the samples. Table 7 provides a summary where the ratings are aggregated for each sample, while Figure 19 visualizes the overall distribution of ratings across all 15 samples, divided by intelligibility and naturalness. The MOS for each sample is calculated by summing the ratings and dividing by 25, the number of participants. Table 8 presents the average intelligibility and naturalness scores for each sample. Thereafter, these values are averaged again

by summing the means of all samples and dividing by 15, the total number of samples. The final results from User Survey 1, representing the average perceived performance of Model 3 on a scale from 1 to 5, where 5 is the best, are as follows:

- Intelligibility MOS: **3.86**
- Naturalness MOS: **3.10**

	Intelligibility					Naturalness				
	Bad	Poor	Fair	Good	Excellent	Bad	Poor	Fair	Good	Excellent
S1	2	11	7	5		3	14	3	3	2
S2		1	4	12	8		7	11	7	
S3		4	9	10	2	3	10	9	3	
S4		2	4	12	7	2	6	8	7	2
S5		1	1	9	14		2	12	8	3
S6		2	2	9	12	1	2	10	10	2
S7		2	4	9	10		2	14	7	2
S8		4	6	12	3		9	12	3	1
S9		1	4	12	8		6	9	9	1
S10		1	3	14	7		3	12	9	1
S11			2	15	8		3	9	11	2
S12	2	9	8	4	2	3	11	10	1	
S13		1	4	8	12		4	8	11	2
S14			2	18	5		3	9	12	1
S15		1	4	14	6		6	11	7	1

Table 7: Summary of the intelligibility and naturalness scores for the samples in User Survey 1. Each row indicates the scores for a sample (S).

	Intelligibility MOS	Naturalness MOS
S1	2.60	2.48
S2	4.08	3.00
S3	3.40	2.48
S4	3.96	3.04
S5	4.44	3.48
S6	4.24	3.40
S7	4.08	3.36
S8	3.56	2.84
S9	4.08	3.20
S10	4.08	3.32
S11	4.24	3.48
S12	2.80	2.36
S13	4.24	3.44
S14	4.12	3.44
S15	4.00	3.12

Table 8: Average intelligibility and naturalness scores for each sample in User Survey 1. Each row indicates the mean score for a sample (S).

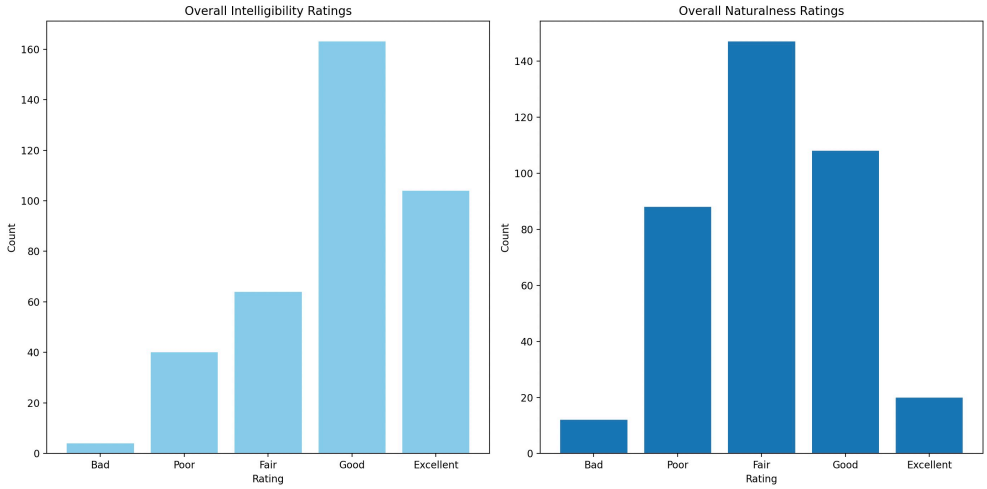

Figure 19: The distribution of intelligibility and naturalness ratings across all 15 samples evaluated in User Survey 1.

5.2.2 User Survey 2: Evaluating Model 4

Figure 51 shows that 16 participants completed User Survey 2. The majority, 13 participants, were in the age group 24–30, while 2 participants were in the age group 50+, and 1 participant was in the age group 18–23. Figures 52, 53, 54, 55, and 56 provide an overview of the participants regarding hearing impairment, Norwegian as their first language, headset usage, noisy environment, and device type used while listening to the speech samples.

Similar to User Survey 1, there was only one participant who did not have Norwegian as their first language. Unlike User Survey 1, one participant in the age group 18–23 reported having a hearing impairment, and another preferred not to answer. As it is unknown whether the person preferring not to answer had a hearing impairment, their response was kept and deemed valid. However, the response from the participant reporting the hearing impairment was deemed invalid due to the potential impact on the rating. Consequently, this response was removed, reducing the total number of valid participants to 15 and removing the age group 18–23 from the statistics entirely.

Moreover, 10 participants reported using headsets while 5 did not, and 13 participants were in a quiet environment while 2 were surrounded by noise. Lastly, 3 participants used a PC, 6 used a Mac, 5 used a phone, and 1 used another device. The majority of participants being in a quiet environment and using headsets is beneficial for the consistency of the answers. However, the diverse range of devices used could potentially impact the results.

15 participants rated the intelligibility and naturalness of 15 different speech samples, resulting in 225 ratings for both intelligibility and naturalness. Table 9 shows the distribution of these ratings for each sample, while Figure 20 visualizes the aggregated distribution of ratings across all samples. The MOS was calculated using the same method as in User Survey 1. Table 10 displays the average intelligibility and naturalness scores for each sample. The final results from User Survey 2, representing the average perceived performance of Model 4 on a scale from 1 to 5, with 5 being the best, are as follows:

- Intelligility MOS: **3.96**
- Naturalness MOS: **3.14**

	Intelligibility					Naturalness				
	Bad	Poor	Fair	Good	Excellent	Bad	Poor	Fair	Good	Excellent
S1		2	7	5	1		7	7	1	
S2		1	2	9	3		5	5	5	
S3		1	7	5	2	1	4	8	2	
S4			1	11	3		4	7	4	
S5			5	7	3		3	9	3	
S6				11	4		1	4	7	3
S7			1	10	4			8	6	1
S8			5	7	3		4	6	4	1
S9				13	2		1	7	6	1
S10			2	10	3			8	6	1
S11			1	11	3		2	7	4	2
S12			6	7	2		5	9	1	
S13			1	10	4		2	8	5	
S14			2	9	4		2	9	4	
S15			2	12	1		4	9	2	

Table 9: Summary of the intelligibility and naturalness scores for the samples in User Survey 2. Each row indicates the scores for a sample (S).

	Intelligibility MOS	Naturalness MOS
S1	3.33	2.60
S2	3.93	3.00
S3	3.53	2.73
S4	4.13	3.00
S5	3.87	3.00
S6	4.27	3.80
S7	4.20	3.53
S8	3.87	3.13
S9	4.13	3.47
S10	4.07	3.53
S11	4.13	3.40
S12	3.73	2.73
S13	4.20	3.20
S14	4.13	3.13
S15	3.93	2.87

Table 10: Average intelligibility and naturalness scores for each sample in User Survey 2. Each row indicates the mean score for a sample (S).

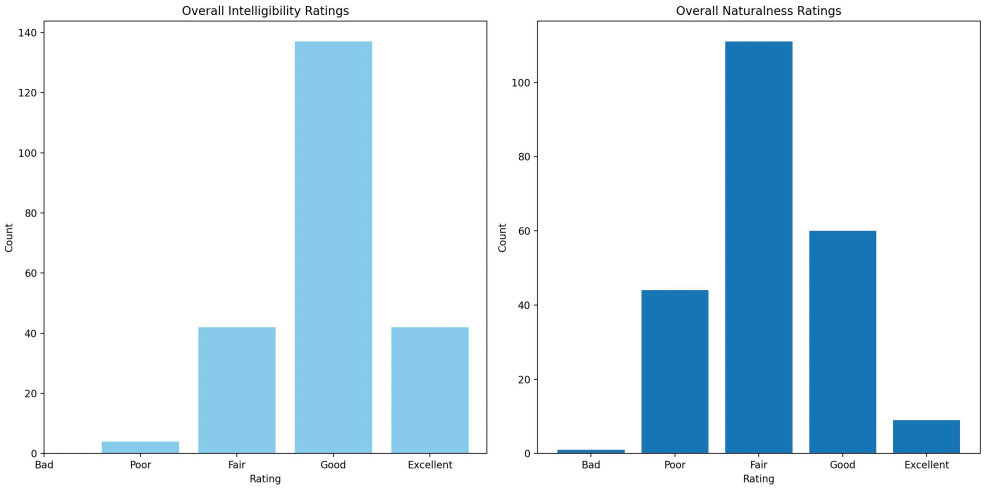

Figure 20: The distribution of intelligibility and naturalness ratings across all 15 samples evaluated in User Survey 2.

5.2.3 User Survey Comparisons

The intelligibility MOS and naturalness MOS showed a slight improvement between the user survey iterations, with scores of **3.86** and **3.10** for User Survey 1, compared to scores of **3.96** and **3.14** for User Survey 2. This indicates that the speech generated by the speaker in Model 4 was perceived as slightly more intelligible and natural than that from the same speaker in Model 3. This aligns with the results from Iteration 2 and Iteration 3 in the objective evaluation, where the intelligibility of the same speakers was assessed by calculating the WER, scoring **17.25%** for Model 3 and **14.95%** for Model 4, showing a similar improvement between iterations. The consistent results across both evaluation methods suggest that the model benefited from combining the datasets, leveraging a larger dataset to enhance training.

Despite User Survey 1 having 10 more participants than User Survey 2, both surveys had relatively similar distributions in terms of age groups, headset usage, noise levels in the environment, and devices used for listening to the speech samples. In both surveys, most responses contributed positively to the consistency and reliability of the data. However, the wide variety of devices used in both surveys posed a potential challenge to uniformity.

At the end of the surveys, after rating all the speech samples, the participants were instructed to select some options about their overall impression of the speech. These opinions are distributed in Figures 21 and 22. The following are the meanings of the different colors, in order:

- Blue - It was difficult to understand the first 1-2 words of the sentences
- Orange - It was difficult to understand the last 1-2 words of the sentences
- Green - It was difficult to understand the middle of the sentences
- Red - It was difficult to understand the whole sentences
- Purple - It was not difficult to understand the sentences
- Brown - It was single words occasionally that were difficult to understand
- Pink - Other

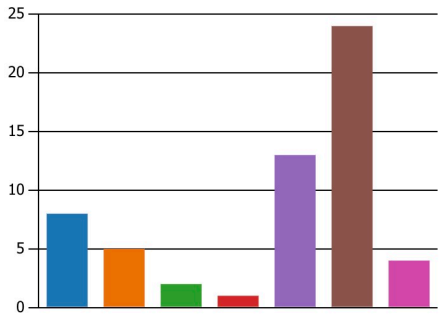

Figure 21: The distribution of overall impression of the speech for User Survey 1.

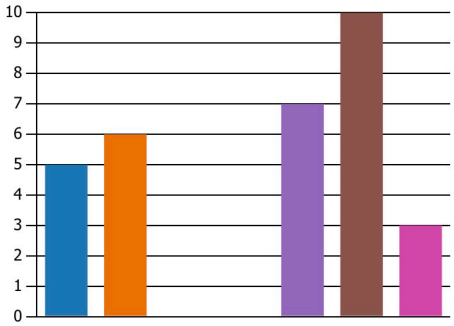

Figure 22: The distribution of overall impression of the speech for User Survey 2.

The general perception is relatively equally balanced between the two user surveys, except for blue and orange switching places. In User Survey 1, participants found it more challenging to understand the first 1-2 words of the sentences, while in User Survey 2, it was reversed, with more participants finding it challenging to understand the last 1-2 words. Although the difference in votes was minor and potentially caused by randomness, it might indicate that Model 4 learned unwanted behavior from Dataset 1 and Dataset 2, making it stop the sentences abruptly. Overall, the feedback indicated that the speech was somewhat difficult to understand at both the beginning and end of sentences, as well as certain words throughout. This issue is likely attributed to inaccuracies in the audio splitting pre-processing step. The built-in silence finder tool in Audacity often cut the audio slightly too early or too late, inadvertently teaching Model 2 and Model 3 to do the same. In hindsight, a more precise approach, such as implementing a Python script to identify and split the audio, could have potentially mitigated this issue.

Further analysis was conducted to evaluate how well the speech samples of varying lengths were rated, providing insight into whether the model performed better at synthesizing speech of different lengths. A Python script was implemented to categorize the 15 samples used in the user surveys based on their word count: short, medium, and long. Samples with a word count at or below 25% of the maximum word count were placed in the short category, those with word counts between 25% and 75% of the maximum word count were placed in the medium category, and those with word counts above 75% of the maximum word count were placed in the long category. These thresholds were selected to ensure the medium category covered the majority of the spectrum. In total, the short category contained 4 samples, the medium category contained 8 samples, and the long category contained 3 samples.

The text file containing the samples was ordered by increasing word count, meaning the first 4 samples of the user surveys were in the short category, the middle 8 were in the medium category, and the last 3 were in the long category. After categorizing the samples, the average scores for each category were calculated and are displayed for User Survey 1 in Figure 23 and for User Survey 2 in Figure 24. These scores are also numerically summarized in Table 11.

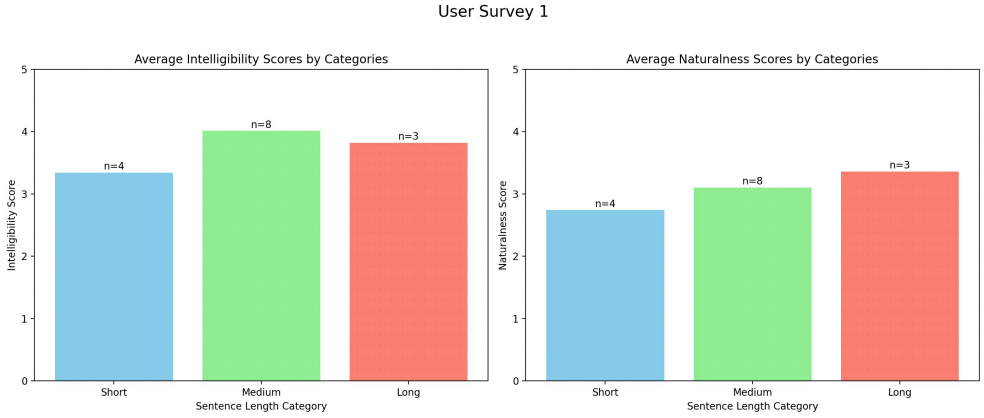

Figure 23: The average intelligibility and naturalness scores by sentence length divided by categories short, medium, and long for User Survey 1.

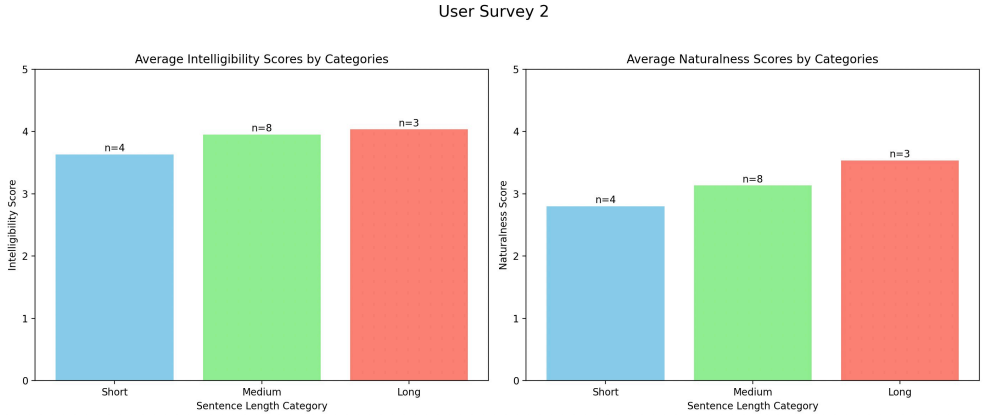

Figure 24: The average intelligibility and naturalness scores by sentence length divided by categories short, medium, and long for User Survey 2.

	User Survey 1		User Survey 2	
	Intelligibility	Naturalness	Intelligibility	Naturalness
Short	3.34	2.74	3.63	2.8
Medium	4.01	3.1	3.95	3.13
Long	3.82	3.36	4.03	3.53

Table 11: Average Intelligibility and Naturalness Scores by Sentence Length Categories (Short, Medium, Long) for User Surveys 1 and 2.

To begin with, all categories showed slightly higher scores in User Survey 2 than in User Survey 1, except for the medium intelligibility category, where User Survey 1 had a marginally higher average score. This aligns with the overall finding that both the intelligibility MOS and naturalness MOS were slightly higher for User Survey 2 than for User Survey 1.

For both surveys, the naturalness and intelligibility scores tended to be lowest for the short category and highest for the long category, with the medium category falling in between. However, the average intelligibility scores for User Survey 1 diverged from this trend, with the medium category scoring the highest and the long category scoring in the middle.

These results generally indicate that the longer the sentence, the better it was rated. Neither Dataset 1, which was used to train Model 1 for User Survey 1, nor Dataset 2 and Dataset 3, which were also used to train Model 3 for User Survey 2, had a heavy distribution of long sentences. This suggests that the perceived quality of longer sentences may have been higher for the participants, rather than the actual quality being better. The finding that medium intelligibility scored higher than long intelligibility in User Survey 1 aligns more with expectations, as it reflects the word count distribution of the dataset used to generate the speech for that survey.

Chapter 6

Discussion

This chapter will discuss the results of the different evaluation methods. Additionally, it will discuss some potential biases regarding the evaluation and data that might have impacted the results to various degrees.

6.1 Objective Evaluation

As demonstrated in Section 5.1, there was a significant variation in intelligibility among the models trained on different datasets. In Iteration 2, when evaluating Models 1, 2, and 3, which were trained on their respective datasets, Model 1 achieved a 21.44% WER, Model 2 got a 31.92% WER, and Model 3 got a 17.25% WER. In Iteration 3, when evaluating Model 4, which was trained on a combination of all three datasets, Speaker 207 recorded a WER of 27.41%, Speaker 1000 achieved 21.67%, and Speaker 1001 scored 14.95%. While these values provide an estimate of intelligibility, they should not be considered definitive due to potential biases.

To begin with, Model 1 was trained on ASR data. This ASR data was derived from the dataset used to train the Norwegian Whisper model, which transcribed the synthetic speech during the objective evaluation. Although the ASR data used in Dataset 1 consisted of only a small portion of the extensive dataset used to train the Whisper model, it might have influenced the results in favor of Model 1.

Second, ASR models are generally trained to recognize speech in noisy environments, anticipating that end users might be surrounded by noise when using a TTS system. Consequently, the data used to train ASR models is often noisy by nature or by injection. Since Model 1 was trained on ASR data, the speech it generated was significantly noisier compared to the other two

models. This conclusion was drawn from analyzing the Mel-spectrograms of the synthetic speech generated by the different models, as discussed in Subsection 4.3.1.4. Given that the Whisper model is designed to operate effectively in noisy environments, it likely accommodated the noise in Model 1 without significantly affecting the WER, potentially resulting in a synthetically low WER in Iteration 2. When Dataset 1 was denoised before training Model 4, the WER increased significantly in Iteration 3. This suggests that the presence of noise had initially masked the true intelligibility issues in Model 1 during Iteration 2, corresponding to Speaker 207 in Iteration 3, further supporting the idea that the ASR model's noise accommodation contributed to the seemingly better performance of Model 1.

Lastly, since Model 1 was trained on a multi-speaker dataset, it can generate speech from all the included speakers. Initial tests during Iteration 1 identified the most intelligible speaker, who also had the most utterances in the dataset. However, this speaker has a different dialect compared to the speakers from the other models. The design of the Norwegian Whisper model, particularly its handling of various dialects, might have influenced the results. According to the paper describing the NPSC dataset, which is part of the training data for the Whisper model and constitutes Dataset 1 in this thesis, the NPSC dataset positively impacted the Norwegian Whisper model by improving the WER score across different dialects [38]. Given this information, along with the relatively good WER score obtained in the objective evaluation, it is reasonable to assume that the model handles dialects well. However, the exact extent of this impact on the results remains uncertain.

6.2 Subjective Evaluation

To begin with, it is important to note that the MOS evaluation is a subjective assessment, meaning the results should be viewed as indicative rather than definitive measures of the model's performance. This subjectivity makes it challenging to directly compare these results with those from similar tests in other studies. However, within the context of this thesis, the MOS evaluation remains valuable for comparing Model 3 and Model 4. Additionally, potential biases may arise from participant-related factors due to budget constraints and a limited social network, which will be discussed in this subsection. Some of these biases will be discussed in this subsection.

When conducting a user survey, it is essential to gather people who represent a small cross-section of the target group. TTS technology becoming ubiquitous has led to a widespread target group, including all people using

technology, which theoretically could span the entire age spectrum. Therefore, an effort to collect samples within the various ranges was made. However, due to a limited social network, it was challenging to recruit a diverse age group, resulting in a clear majority of participants being in the 24–30 years age range in both iterations of the user surveys. The MOS scores might have differed with a more varied age distribution, as perceptions of speech can vary across age groups. Although using paid user survey services to obtain a more randomized sample could have been an alternative, budget and time constraints made this option unfeasible. Consequently, most participants in both user surveys were known to varying degrees, which could have influenced the results by potentially elevating the MOS scores. People may subconsciously provide more favorable feedback to those they know. While this is difficult to quantify, it is an important factor to consider in the evaluation.

Conducting the user surveys in two separate iterations introduces a few variables. Firstly, it is unclear whether the participants were the same individuals for both surveys. Different participants can have different perceptions of speech, which could affect their ratings. Additionally, if a participant took both surveys, they might have rushed through the second time due to familiarity, potentially not listening as carefully and rating less thoroughly. Secondly, User Survey 1 had 10 more participants than User Survey 2, meaning the first survey had a larger data set to base the MOS on compared to the second survey. These discrepancies between the user surveys could have impacted the MOS to some extent.

Another important aspect to consider is that the speech samples evaluated in the user surveys were limited to sentences ending with periods, mirroring the data on which the models were trained. As a result, the MOS does not account for expressiveness or stress, such as changes in intonation for exclamatory sentences. Consequently, the naturalness MOS would likely be lower if a variety of sentence types had been included in the user surveys. This limitation makes the model a PoC, as it would require coverage for all types of sentences to be applicable in real-world settings.

Lastly, as previously noted, there was inconsistency in participants' use of headsets, devices, and environments, including noisy settings. Despite efforts to maintain consistency in responses, these variations may have influenced the results. For instance, speech may sound different through headsets compared to computer speakers. This lack of uniformity in listening conditions could have affected participants' perceptions and ratings, thereby impacting the overall MOS results.

6.3 Data and Model Training

Similarly to the evaluation process, certain aspects of data collection, pre-processing, and model training may have influenced the performance of the models. This section will delve into these factors and present additional observations.

6.3.1 Audio Duration Distribution

As depicted in Figure 6, Dataset 2 exhibits a very short audio duration distribution. The minimum utterance length is under 1 second, the maximum is 7 seconds, and the average is 1.94 seconds. As detailed in Section 2.3.1, a well-rounded dataset with a mix of short and long utterances is crucial for a model to learn effectively, especially for edge cases. Consequently, Model 2 may have struggled with longer sentences, potentially impacting its overall performance. In hindsight, subscribing to an audiobook service, even on a trial basis, could have provided access to a larger and more varied dataset, potentially enhancing the model's training and performance.

6.3.2 Whisper Transcription

Another factor that may have adversely affected the overall performance of all the models was the exclusive reliance on automatic transcription using Whisper for Dataset 2 and Dataset 3, without manual verification of transcription quality. While this approach significantly saved time and effort, the Norwegian Whisper model [36] has a WER of 8.3%, implying that roughly one in every ten words is incorrectly transcribed.[1] In practice, incorrect transcriptions result in the misalignment of spoken words with their written counterparts, leading to the model learning incorrect pronunciations. This likely impacted the model's performance to some extent.

6.3.3 Mel-spectrogram Mean and Standard Deviation

Another interesting aspect to consider is the comparison of the Mel-spectrogram means and standard deviations for the different datasets. Table 12 presents these values, derived from Table 3. Based solely on these metrics and the TTS dataset characteristics, we can theoretically determine which dataset

[1] https://huggingface.co/NbAiLab/nb-whisper-small-beta

is best suited for TTS development, although other factors, such as dataset size, significantly influence actual performance.

The Mel-spectrogram mean represents the average value of all Mel-spectrograms generated for the entire dataset. A higher mean value indicates a narrower range of frequencies is captured, while a lower mean value suggests a broader range of frequencies. Conversely, the Mel-spectrogram standard deviation measures the variation across the dataset, indicating how much the values deviate from the mean. A higher standard deviation implies greater variability in the speech, while a lower value denotes less variability and more homogeneous data.

Ideally, a lower Mel-spectrogram mean and a higher standard deviation, within reasonable limits, are preferred, as they indicate a more diverse dataset with greater variability, potentially capturing a broader range of speech patterns. However, the size of the dataset and the training process play crucial roles in the final outcome. A sufficiently large dataset is necessary for the model to effectively learn patterns. Therefore, a balance must be struck, as a lower standard deviation may simplify the training process, but it could result in a model that is less adaptable to different speaking styles.

Name	Dataset 1	Dataset 2	Dataset 3
Mel Spectogram Mean	≈ -5.132	≈ -5.681	≈ -7.267
Mel Spectogram Standard Deviation	≈ 1.636	≈ 2.487	≈ 1.878

Table 12: Mel-spectrogram mean and standard deviation values for the three datasets.

Based on the Mel-spectrogram mean values of these datasets, Dataset 3 has a significantly lower mean than Dataset 1 and Dataset 2, indicating a broader range of captured frequencies. Additionally, Dataset 2 exhibits a significantly higher Mel-spectrogram standard deviation compared to Dataset 1 and Dataset 3, suggesting a more varied dataset. After superficially evaluating these values without considering the size of the datasets or other factors, one might conclude that Dataset 3 is the best suited for TTS development, followed by Dataset 2, and lastly, Dataset 1. This conclusion is based on the datasets' capacity to capture diverse patterns, potentially leading to the generation of more varied yet stable speech.

However, it is noteworthy that these observations do not align with the intelligibility performance obtained from Iteration 2 of the objective

evaluation. This discrepancy suggests that, for a low-resource language, the quantity of data may be more critical in training an intelligible TTS model than the quality of the data in terms of variability and range of frequencies. This finding appears to contrast with the conclusions drawn for different speakers within a dataset in Iteration 1. The key difference is that the latter was measured for individual speakers within a multi-speaker dataset rather than the dataset as a whole.

6.3.4 Additional Pre-processing Techniques

The pre-processing steps for Models 1, 2, and 3, including audio splitting, sample rate conversion, and formatting, were selected to meet the minimum requirements for training the model to generate synthetic speech. While these steps were sufficient for basic training, additional pre-processing measures could have been implemented to enhance the models' performance. Two key steps that could have been taken are noise removal to improve data quality and data augmentation to increase the dataset size.

Noise removal was applied to Dataset 1 before training Model 4 to reduce background noise and improve intelligibility. However, the results from Iteration 3 of the objective evaluation indicate that the noise removal process was excessive, resulting in lower voice volume for the speakers in Dataset 1 and decreased intelligibility. This highlights the importance of carefully balancing noise reduction to avoid compromising the fundamental structure of the speech.

Data augmentation, which involves synthetically generating new data from existing data by modifying parameters such as speaking rate and pitch, is commonly used in low-resource language settings due to the limited availability of data. Despite its potential benefits, data augmentation was not utilized in this thesis due to the potential of developing a reasonably performing PoC without relying on extensive pre-processing techniques. Additionally, time constraints limited the feasibility of applying comprehensive data augmentation. However, for future research aimed at creating a more professional artifact or a higher-performing PoC, incorporating data augmentation and careful noise removal could significantly enhance the model's performance.

6.3.5 Training Duration

The four models were trained for varying durations to create the best possible Norwegian TTS model. The iterative development process involved increasing training time for each model to enhance performance. However, Model 4's training time was reduced due to time constraints. This variation in training times makes it challenging to assess the impact on results accurately. If all models had been trained for the same duration, performance differences would solely reflect the quality of the data, not the training time. In other words, the goal of the thesis to create the best possible TTS model conflicted with the need for reliable testing, as the varying training times introduced an additional variable.

Chapter 7

Conclusion and Future Work

In this thesis, we successfully developed a proof-of-concept (PoC) Norwegian text-to-speech (TTS) model, addressing the scarcity of both open-source Norwegian TTS models and Norwegian TTS datasets. Our research focused on identifying the most suitable speech data for TTS development in low-resource languages and evaluating the naturalness and intelligibility of the generated synthetic speech.

We decided to develop a proof of concept (PoC) adult speech TTS model as an initial step towards creating an open-source adult speech TTS model, and eventually, a child speech TTS model. To create an effective child speech TTS model, it is essential to utilize transfer learning techniques with a well-developed adult speech TTS model as the foundation. Child speech TTS is crucial for applications such as realistic child avatars, which can be used for police interview training to reduce child abuse.

To develop a PoC adult speech TTS model, we collected and prepared three distinct datasets: nearly three hours of automatic speech recognition (ASR) data from the Norwegian Parliamentary Speech Corpus (NPSC), nearly one hour of audiobook data, and nearly two hours of self-recorded speech data. Each dataset underwent pre-processing steps to ensure compatibility with the Matcha-TTS architecture used for training our models. Four models were developed: Model 1 trained on the ASR data, Model 2 trained on audiobook data, Model 3 trained on self-recorded data, and Model 4 trained on a combination of all three datasets. The models were evaluated using both objective metrics, specifically Word Error Rate (WER), and subjective metrics, specifically Mean Opinion Score (MOS), providing insights into the strengths and weaknesses of each dataset type.

The primary research questions addressed were: *What type of speech data*

is most suitable for TTS development in a low-resource language? and *How do the models perform in terms of naturalness and intelligibility?* Our findings indicate that while all datasets had merits, self-recorded data offered a balance between quality and practicality, yielding relatively higher intelligibility and naturalness scores. However, combining ASR and audiobook datasets also showed potential.

The objective testing was split into iterations: initially, Models 1, 2, and 3 trained on individual datasets were evaluated, followed by testing the same speakers from Model 4 trained on the combined dataset. Model 1 achieved a WER of 21.44%, Model 2 had a WER of 31.92%, and Model 3 had a WER of 17.25%. Model 4, which combined all three datasets making it a multi-speaker dataset, demonstrated improved performance for two of the three models with Speaker 207, representing Model 1, achieving a WER of 27.41%; Speaker 1000, representing Model 2, achieving a WER of 21.67%; and Speaker 1001, representing Model 3, achieving a WER of 14.95%. These results indicate that combining datasets can enhance model performance.

In the subjective evaluation, the MOS for intelligibility and naturalness showed a slight improvement between User Survey 1 and User Survey 2. The user surveys tested the same speakers, with User Survey 1 evaluating Model 3 and User Survey 2 evaluating Model 4. For User Survey 1, the MOS for intelligibility was 3.86 and for naturalness was 3.10, indicating that decent synthetic speech can be generated from just two hours of training data. For User Survey 2, the MOS for intelligibility was 3.96 and for naturalness was 3.14, demonstrating that an additional four hours of training data can enhance performance. These results suggest that the speech generated by Model 4 was perceived as slightly more intelligible and natural than that of Model 3.

Our iterative approach facilitated comparisons between models and in the subjective evaluation revealing that the longer the sentence, the better it was rated, despite the datasets not having a heavy distribution of long sentences. This suggests that perceived quality may be influenced by factors beyond dataset characteristics. Additionally, the subjective nature of MOS evaluations, introducing some potential biases for the survey, should be considered when interpreting the results.

Throughout this thesis, several recommendations regarding datasets have emerged. It was observed that in large enough multi-speaker datasets, the quality and distribution of sentence lengths for each speaker are more critical than the sheer amount of data, as the model can learn basic patterns from other speakers in the dataset. For datasets as a whole, the overall volume of data proved to be crucial for training effective TTS models, as an insufficient

amount would not provide the model with enough patterns to learn from. This finding underscores the importance of utilizing all available data, especially in low-resource languages. Researchers and developers should not hesitate to start with ”imperfect” or ”bad” data, as it can still significantly contribute to the development of a functional TTS model and enhance the model's ability to generalize better.

The foundations laid by this thesis can have profound social and sustainability impacts. Norwegian TTS technology has the potential to significantly aid individuals with disabilities, such as those with visual impairments, by improving accessibility to information and services. Enhanced TTS can also elevate user experiences across various systems, including GPS, virtual assistants, and audiobooks, making technology more user-friendly and inclusive. Moreover, this research contributes to sustainability by facilitating the development of advanced educational tools and fostering better learning experiences for future generations.

Future work should focus on achieving the ultimate goal of developing a child speech TTS model by expanding and diversifying datasets through techniques such as data augmentation and carefully applying noise reduction, ensuring the voice volume is preserved. Additionally, developing adult speech models capable of handling various sentence types and intonations is essential. Conducting more controlled and diverse user surveys will provide more reliable subjective evaluations. Ultimately, the insights gained from this research will guide future advancements in TTS technology, contributing to the creation of high-quality synthetic speech for various applications.

References

[1] C. Widom, "Longterm Consequences of Child Maltreatment," *Handbook of child maltreatment*, vol. 2, pp. 225–247, Nov. 2014, ISSN: 978-94-007-7207-6. DOI: 10.1007/978-94-007-7208-3_12.

[2] Louise Dixon, Daniel F. Perkins, Cathrine Hamilton-Giachritsis, and Leam A. Craig, "The Wiley Handbook of What Works in Child Maltreatment: An Evidence-Based Approach to Assessment and Intervention in Child Protection," en, in *The Wiley Handbook of What Works in Child Maltreatment*, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/97811189761 John Wiley & Sons, Ltd, 2017, pp. i–xxv, ISBN: 978-1-118-97611-1. DOI: 10.1002/9781118976111.fmatter. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118976111.fmatter (visited on 04/01/2024).

[3] J. A. Adams, K. J. Farst, and N. D. Kellogg, "Interpretation of Medical Findings in Suspected Child Sexual Abuse: An Update for 2018," eng, *Journal of Pediatric and Adolescent Gynecology*, vol. 31, no. 3, pp. 225–231, Jun. 2018, ISSN: 1873-4332. DOI: 10.1016/j.jpag.2017.12.011.

[4] D. A. Brown and M. E. Lamb, "Forks in the road, routes chosen, and journeys that beckon: A selective review of scholarship on children's testimony," en, *Applied Cognitive Psychology*, vol. 33, no. 4, pp. 480–488, 2019, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3511, ISSN: 1099-0720. DOI: 10.1002/acp.3511. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3511 (visited on 04/02/2024).

[5] M. Johnson, S. Magnussen, C. Thoresen, K. Lønnum, L. V. Burrell, and A. Melinder, "Best Practice Recommendations Still Fail to Result in Action: A National 10-Year Follow-up Study of Investigative Interviews in CSA Cases," en, *Applied Cognitive Psychology*, vol. 29, no. 5, pp. 661–668, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3147,

ISSN: 1099-0720. DOI: 10.1002/acp.3147. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3147 (visited on 04/07/2024).

[6] G. A. Baugerud *et al.*, "Multimodal Virtual Avatars for Investigative Interviews with Children," en, in *Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, Taipei Taiwan: ACM, Aug. 2021, pp. 2–8, ISBN: 978-1-4503-8529-9. DOI: 10.1145/3463944.3469269. [Online]. Available: https://dl.acm.org/doi/10.1145/3463944.3469269 (visited on 03/14/2024).

[7] P. Salehi *et al.*, "Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children," en, *Big Data and Cognitive Computing*, vol. 6, no. 2, p. 62, Jun. 2022, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2504-2289. DOI: 10.3390/bdcc6020062. [Online]. Available: https://www.mdpi.com/2504-2289/6/2/62 (visited on 03/31/2024).

[8] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, *A Survey on Neural Speech Synthesis*, arXiv:2106.15561 [cs, eess], Jul. 2021. DOI: 10.48550/arXiv.2106.15561. [Online]. Available: http://arxiv.org/abs/2106.15561 (visited on 04/10/2024).

[9] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, *Matcha-TTS: A fast TTS architecture with conditional flow matching*, arXiv:2309.03199 [cs, eess], Jan. 2024. DOI: 10.48550/arXiv.2309.03199. [Online]. Available: http://arxiv.org/abs/2309.03199 (visited on 03/11/2024).

[10] E. Cooper, E. Li, and J. Hirschberg, "Characteristics of Text-to-Speech and Other Corpora," 2018, pp. 690–694. DOI: 10.21437/SpeechProsody.2018-140. [Online]. Available: https://www.isca-archive.org/speechprosody_2018/cooper18b_speechprosody.html (visited on 03/21/2024).

[11] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. Clark, and J. Yamagishi, "TUNDRA: A multilingual corpus of found data for TTS research created with light supervision," Aug. 2013. DOI: 10.21437/Interspeech.2013-545.

[12] A. Gallardo Antolín, J. M. Montero, and S. King, *A Comparison of Open-Source Segmentation Architectures for Dealing with Imperfect Data from the Media in Speech Synthesis*, eng. International Speech Communication Association, 2014, ISBN: 978-1-63439-435-2. [Online]. Available: https://hdl.handle.net/10016/21478 (visited on 06/01/2024).

[13] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, *Training Multi-Speaker Neural Text-to-Speech Systems using Speaker-Imbalanced Speech Corpora*, arXiv:1904.00771 [cs, eess, stat], Apr. 2019. DOI: 10.48550/arXiv.1904.00771. [Online]. Available: http://arxiv.org/abs/1904.00771 (visited on 06/01/2024).

[14] E. L. Cooper, "Text-to-Speech Synthesis Using Found Data for Low-Resource Languages," [object Object], 2019. DOI: 10.7916/D8-VDZP-J870. [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/d8-vdzp-j870 (visited on 04/10/2024).

[15] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance Selection for Optimizing Intelligibility of TTS Voices Trained on ASR Data," Aug. 2017, pp. 3971–3975. DOI: 10.21437/Interspeech.2017-465.

[16] M. Y. Yiwere, A. Barcovschi, R. Jain, H. Cucu, and P. Corcoran, "Augmentation Techniques for Adult-Speech to Generate Child-Like Speech Data Samples at Scale," *IEEE Access*, vol. 11, pp. 109 066–109 081, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3317360. [Online]. Available: https://ieeexplore.ieee.org/document/10256249/ (visited on 03/08/2024).

[17] P. Wagner *et al.*, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," en, in *10th ISCA Workshop on Speech Synthesis (SSW 10)*, ISCA, Sep. 2019, pp. 105–110. DOI: 10.21437/SSW.2019-19. [Online]. Available: https://www.isca-archive.org/ssw_2019/wagner19_ssw.html (visited on 03/15/2024).

[18] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," en, in *12th ISCA Speech Synthesis Workshop (SSW2023)*, ISCA, Aug. 2023, pp. 41–47. DOI: 10.21437/SSW.2023-7. [Online]. Available: https://www.isca-ar

chive.org/ssw_2023/kirkland23_ssw.html (visited on 04/27/2024).

[19] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," 1997, pp. 473–476. DOI: 10.21437/Eurospeech.1997-161. [Online]. Available: https://www.isca-archive.org/eurospeech_1997/lee97b_eurospeech.html (visited on 04/05/2024).

[20] Sinno Jialin Pan and Qiang Yang, *A Survey on Transfer Learning | IEEE Journals & Magazine | IEEE Xplore*, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5288526 (visited on 04/02/2024).

[21] N. Kaur and P. Singh, "Conventional and contemporary approaches used in text to speech synthesis: A review," en, *Artificial Intelligence Review*, vol. 56, no. 7, pp. 5837–5880, Jul. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-022-10315-0. [Online]. Available: https://doi.org/10.1007/s10462-022-10315-0 (visited on 03/30/2024).

[22] A. W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, ISSN: 2379-190X, vol. 4, Apr. 2007, pp. IV–1229–IV–1232. DOI: 10.1109/ICASSP.2007.367298. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4218329 (visited on 03/26/2024).

[23] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, ISSN: 2379-190X, May 2013, pp. 7962–7966. DOI: 10.1109/ICASSP.2013.6639215. [Online]. Available: https://ieeexplore.ieee.org/document/6639215 (visited on 03/30/2024).

[24] Y. Wang *et al.*, *Tacotron: Towards End-to-End Speech Synthesis*, arXiv:1703.10135 [cs], Apr. 2017. DOI: 10.48550/arXiv.1703.10135. [Online]. Available: http://arxiv.org/abs/1703.10135 (visited on 03/30/2024).

[25] A. v. d. Oord *et al.*, *WaveNet: A Generative Model for Raw Audio*, arXiv:1609.03499 [cs], Sep. 2016. DOI: 10.48550/arXiv.1609.03499. [Online]. Available: http://arxiv.org/abs/1609.03499 (visited on 03/30/2024).

[26] J. Shen *et al.*, *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*, arXiv:1712.05884 [cs], Feb. 2018. DOI: 10.48550/arXiv.1712.05884. [Online]. Available: http://arxiv.org/abs/1712.05884 (visited on 04/07/2024).

[27] Y. Ren *et al.*, *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*, arXiv:2006.04558 [cs, eess], Aug. 2022. DOI: 10.48550/arXiv.2006.04558. [Online]. Available: http://arxiv.org/abs/2006.04558 (visited on 04/07/2024).

[28] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ISSN: 2379-190X, May 2020, pp. 7209–7213. DOI: 10.1109/ICASSP40776.2020.9054484. [Online]. Available: https://ieeexplore.ieee.org/document/9054484?denied= (visited on 04/07/2024).

[29] J. Kim, S. Kim, J. Kong, and S. Yoon, *Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search*, arXiv:2005.11129 [cs, eess], Oct. 2020. DOI: 10.48550/arXiv.2005.11129. [Online]. Available: http://arxiv.org/abs/2005.11129 (visited on 04/07/2024).

[30] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, *Neural Speech Synthesis with Transformer Network*, arXiv:1809.08895 [cs], Jan. 2019. DOI: 10.48550/arXiv.1809.08895. [Online]. Available: http://arxiv.org/abs/1809.08895 (visited on 04/07/2024).

[31] J. Nouza, L. Mateju, P. Cerva, and J. Zdansky, "Developing State-of-the-Art End-to-End ASR for Norwegian," en, in *Text, Speech, and Dialogue*, K. Ekštein, F. Pártl, and M. Konopík, Eds., Cham: Springer Nature Switzerland, 2023, pp. 200–213, ISBN: 978-3-031-40498-6. DOI: 10.1007/978-3-031-40498-6_18.

[32] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 42, Dec. 2021, ISSN: 1687-4722. DOI: 10.1186/s13636-021-00225-4. [Online]. Available: https://doi.org/10.1186/s13636-021-00225-4 (visited on 03/13/2024).

[33] C. Terblanche, M. Harty, M. Pascoe, and B. V. Tucker, "A Situational Analysis of Current Speech-Synthesis Systems for Child Voices: A Scoping Review of Qualitative and Quantitative Evidence," en, *Applied Sciences*, vol. 12, no. 11, p. 5623, Jan. 2022, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app12115623. [Online]. Available: https://www.mdpi.com/2076-3417/12/11/5623 (visited on 03/14/2024).

[34] M. E. N. Begnum, D. Meen, and T. Nordgård, "A Strategy for Producing High-Quality Norwegian Synthetic Child Voices," en, 2012.

[35] R. Jain, M. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, *A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis*, arXiv:2203.11562 [cs, eess], Apr. 2022. DOI: 10.48550/arXiv.2203.11562. [Online]. Available: http://arxiv.org/abs/2203.11562 (visited on 02/29/2024).

[36] P. E. Kummervold, J. de la Rosa, F. Wetjen, R.-A. Braaten, and P. E. Solberg, *Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges*, en, arXiv:2402.01917 [cs], Feb. 2024. [Online]. Available: http://arxiv.org/abs/2402.01917 (visited on 06/04/2024).

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust Speech Recognition via Large-Scale Weak Supervision*, arXiv:2212.04356 [cs, eess], Dec. 2022. DOI: 10.48550/arXiv.2212.04356. [Online]. Available: http://arxiv.org/abs/2212.04356 (visited on 05/08/2024).

[38] P. E. Solberg and P. Ortiz, *The Norwegian Parliamentary Speech Corpus*, arXiv:2201.10881 [cs, eess], Jan. 2022. DOI: 10.48550/arXiv.2201.10881. [Online]. Available: http://arxiv.org/abs/2201.10881 (visited on 05/06/2024).

[39] *P.800.1 : Mean opinion score (MOS) terminology*, 2016. [Online]. Available: https://www.itu.int/rec/T-REC-P.800.1-201607-I/en (visited on 05/03/2024).

Appendix A

User Survey 1 Results

All the graphs that were generated to show the results of the user survey by Microsoft Forms are shown in this appendix. There were 25 participants in total, and all questions were mandatory. The graphs will be shown in the same order as they were asked, beginning with question 1. As the questions were asked in Norwegian, all figures will have an English translation of the text in the figure description below the image. If a word has already been translated in a previous figure, such as ”Yes” or ”No”, it will not be translated again.

Figure 25: Question 1: ”How old are you?”

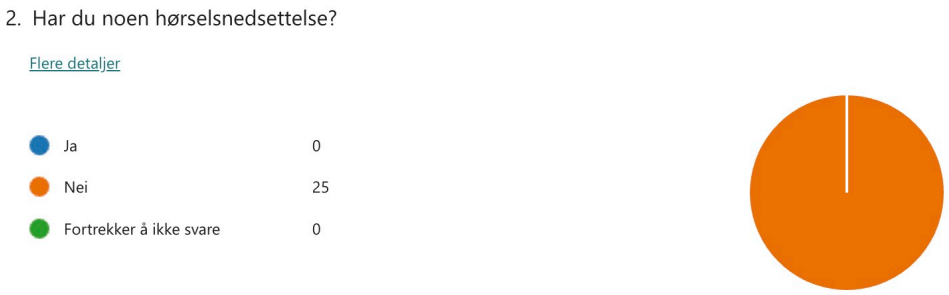

Figure 26: Question 2: ”Do you have any hearing impairment?” Ja=Yes, Nei=No, Foretrekker å ikke svare=Prefer not to answer

Figure 27: Question 3: ”Is Norwegian your first language?”

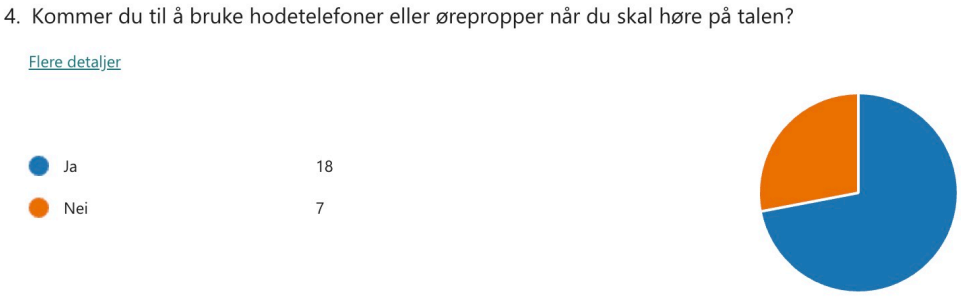

Figure 28: Question 4: ”Will you be using headphones when listening to the speech?”

5. Befinner du deg i et bråkete rom eller er du omgitt av bråk mens du tar testen?

Flere detaljer

Ja	3
Nei	22

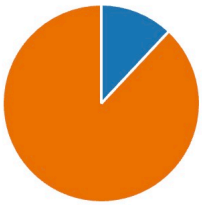

Figure 29: Question 5: ”Are you located in a noisy room or surrounded by noise while taking the test?”

6. Hva slags enhet utfører du evalueringen på?

Flere detaljer

Pc	3
Mac	11
Mobil	11
Nettbrett	0
Annet	0

Figure 30: Question 6: ”What kind of device are you performing the evaluation on?” Mobil=Phone, Nettbrett=Tablet, Annet=Other

7. Hvor ofte bruker du tekst-til-tale (TTT)-teknologi, slik som virtuelle assistenter (for eksempel Siri, Google Assistant) eller GPS-navigasjonssystemer?

Flere detaljer

Aldri	3
Sjeldent	18
Måndelig	2
Ukentlig	0
Daglig	2

Figure 31: Question 7: ”How often do you use text-to-speech (TTS)-technology, such as virtual assistants (e.g. Siri, Google Assistant) or GPS-navigation systems?” Aldri=Never, Sjeldent=Rarely, Månedlig=Monthly, Ukentlig=Weekly, Daglig=Daily

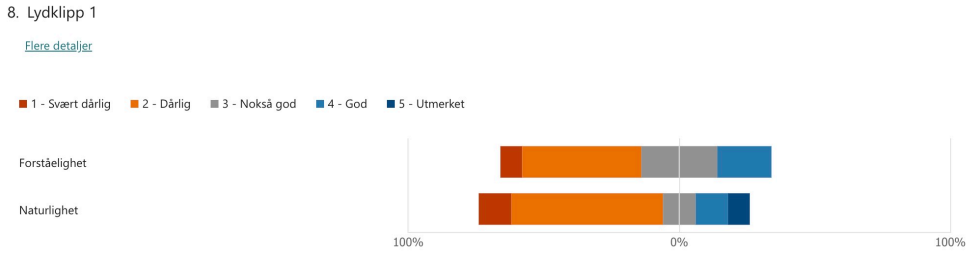

Figure 32: Question 8: ”Audio clip 1” Forståelighet=Intelligibility, Naturlighet=Naturalness, Svært dårlig=Bad, Dårlig=Poor, Nokså god=Fair, God=Good, Utmerket=Excellent

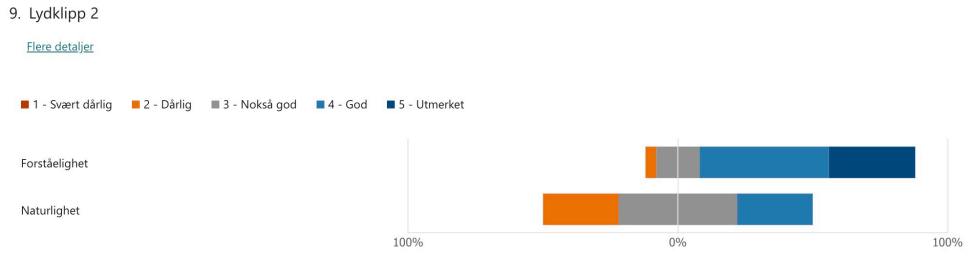

Figure 33: Question 9: ”Audio clip 2”

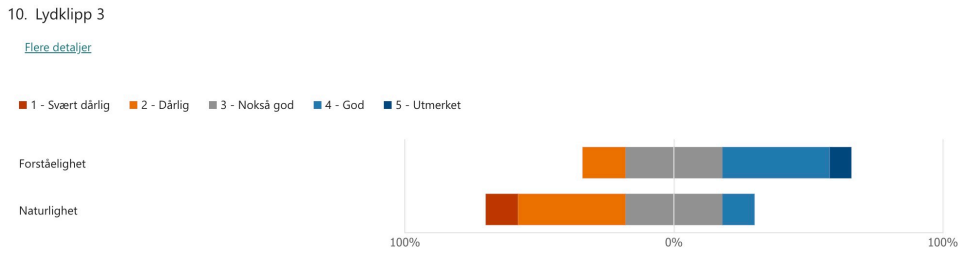

Figure 34: Question 10: ”Audio clip 3”

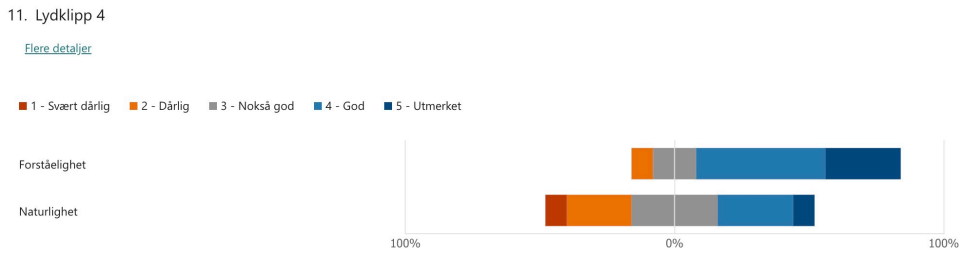

Figure 35: Question 11: ”Audio clip 4”

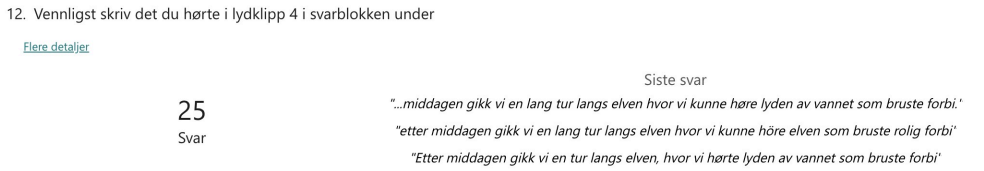

Figure 36: Question 12: ”Please write what you heard in audio clip 4 in the text block below”

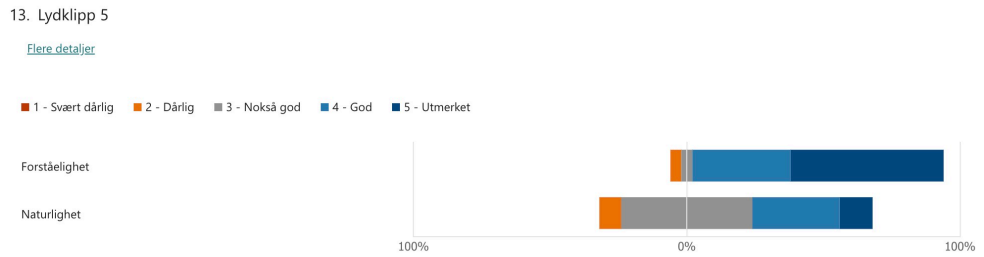

Figure 37: Question 13: ”Audio clip 5”

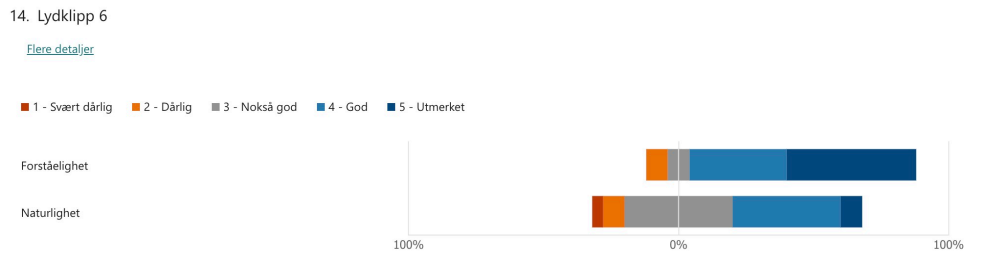

Figure 38: Question 14: ”Audio clip 6”

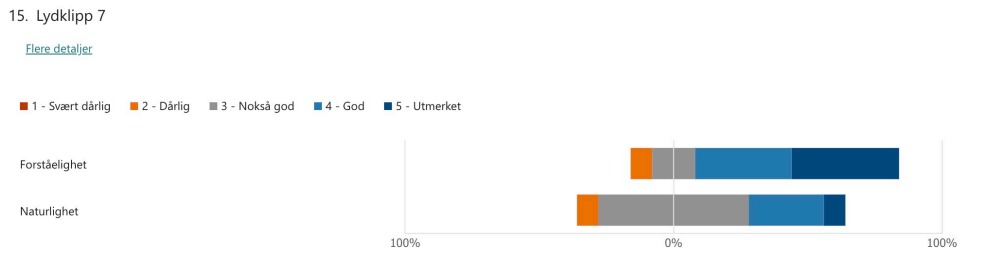

Figure 39: Question 15: ”Audio clip 7”

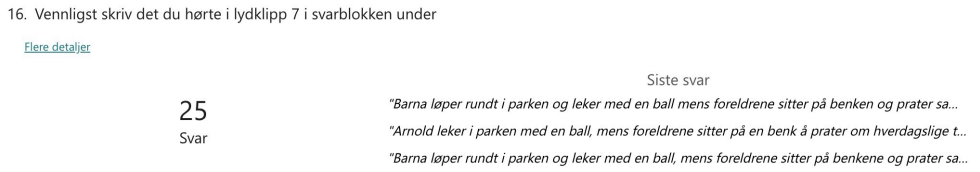

Figure 40: Question 16: ”Please write what you heard in audio clip 7 in the text block below”

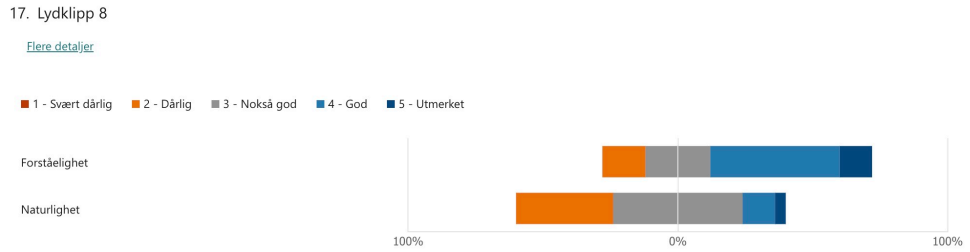

Figure 41: Question 17: ”Audio clip 8”

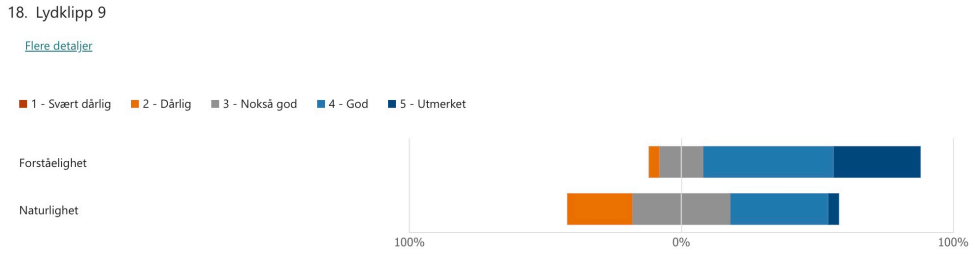

Figure 42: Question 18: ”Audio clip 9”

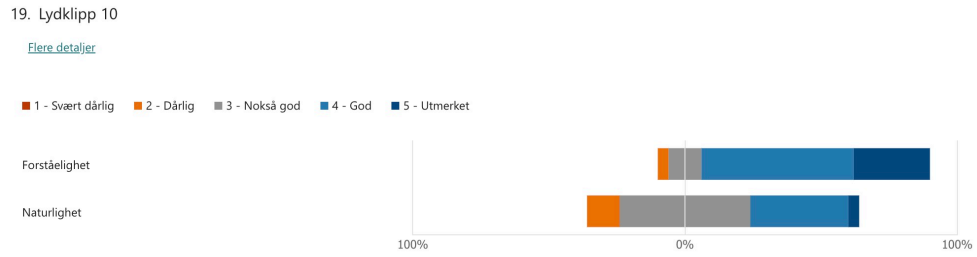

Figure 43: Question 19: ”Audio clip 10”

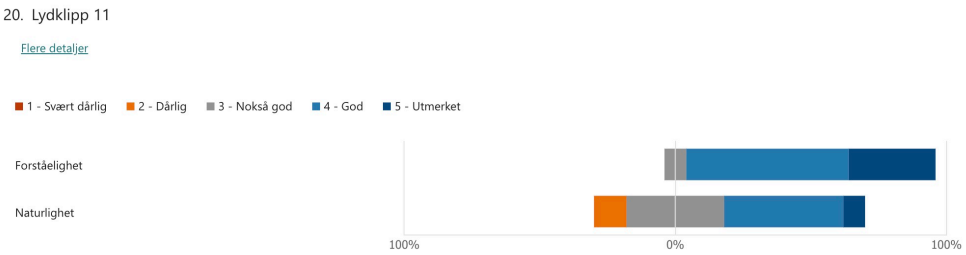

Figure 44: Question 20: ”Audio clip 11”

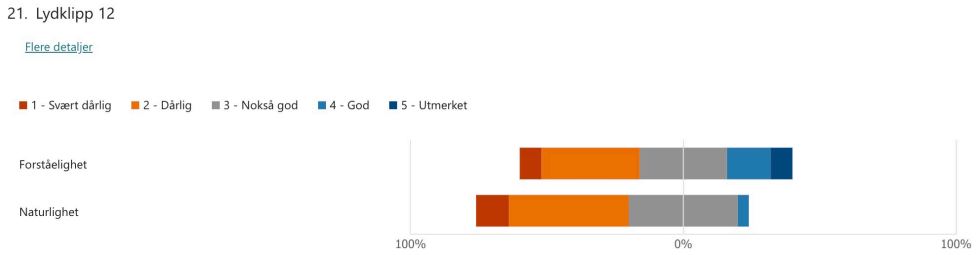

Figure 45: Question 21: ”Audio clip 12”

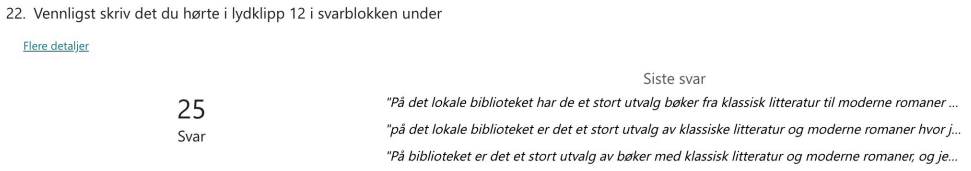

Figure 46: Question 22: ”Please write what you heard in audio clip 12 in the text block below”

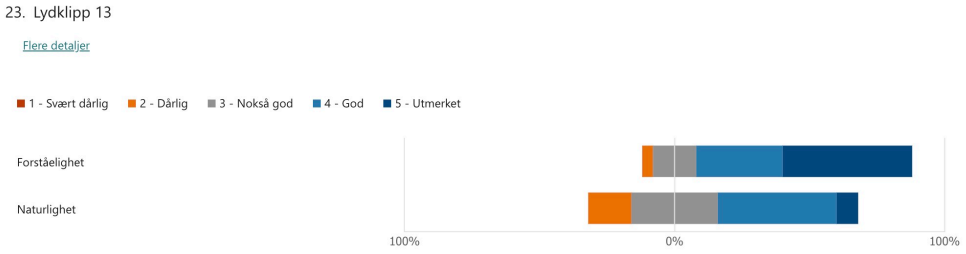

Figure 47: Question 23: ”Audio clip 13”

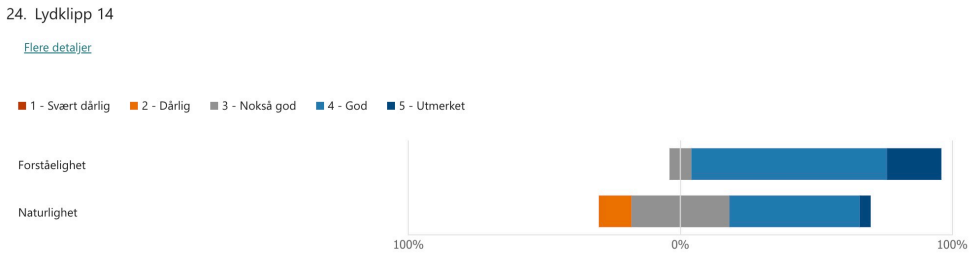

Figure 48: Question 24: ”Audio clip 14”

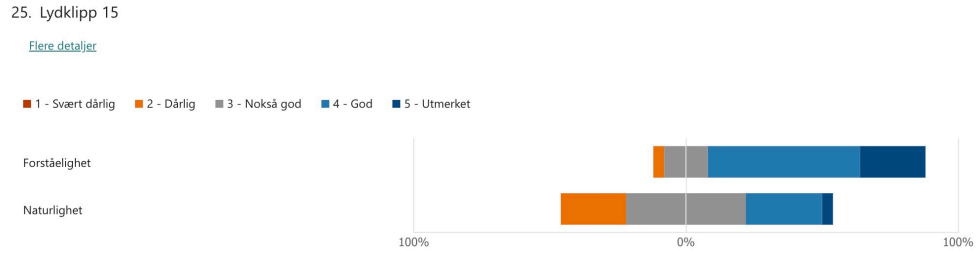

Figure 49: Question 25: ”Audio clip 15”

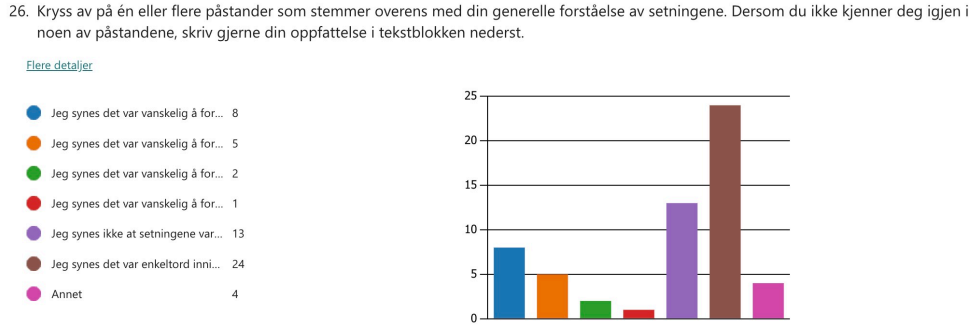

Figure 50: Question 26: ”Select one or more statements that match your general understanding of the sentences. If none of the statements apply to you, please write your interpretation in the text block below.” Blue=It was difficult to understand the first 1-2 words of the sentence, Orange=It was difficult to understand the last 1-2 words of the sentences, Green=It was difficult to understand the middle of the sentences, Red=It was difficult to understand the whole sentences, Purple=It was not difficult to understand the sentences, Brown=It was single words occasionally that were difficult to understand, Pink=Other

Appendix B

User Survey 2 Results

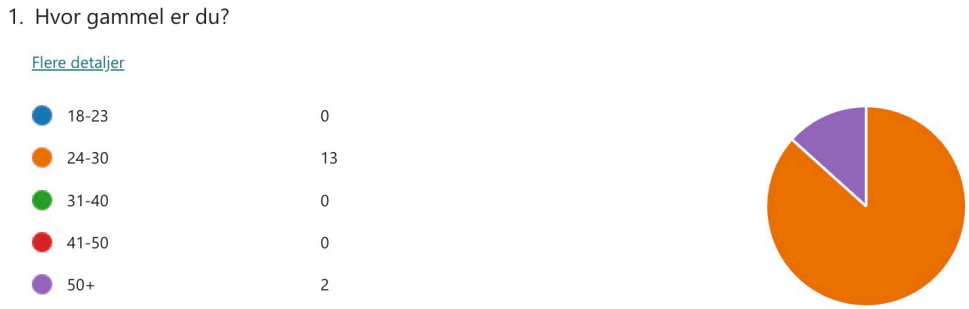

Figure 51: Question 1: ”How old are you?”

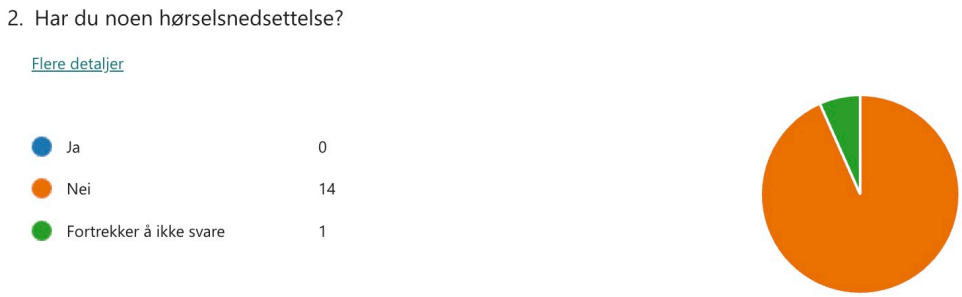

Figure 52: Question 2: ”Do you have any hearing impairment?” Ja=Yes, Nei=No, Foretrekker å ikke svare=Prefer not to answer

3. Er norsk ditt morsmål?

Flere detaljer

Ja	14
Nei	1

Figure 53: Question 3: ”Is Norwegian your first language?”

4. Kommer du til å bruke hodetelefoner eller øreproppper når du skal høre på talen?

Flere detaljer

Ja	10
Nei	5

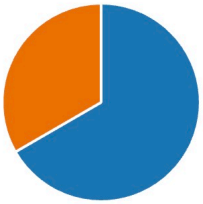

Figure 54: Question 4: ”Will you be using headphones when listening to the speech?”

5. Befinner du deg i et bråkete rom eller er du omgitt av bråk mens du tar testen?

Flere detaljer

Ja	2
Nei	13

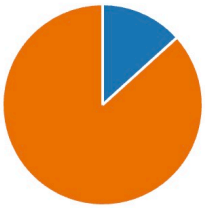

Figure 55: Question 5: ”Are you located in a noisy room or surrounded by noise while taking the test?”

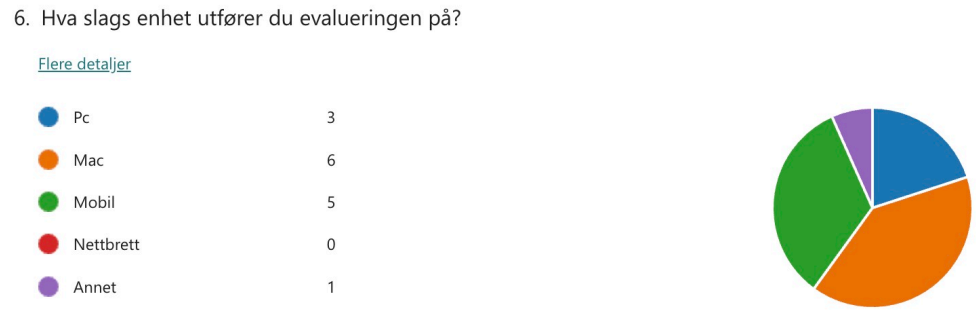

Figure 56: Question 6: "What kind of device are you performing the evaluation on?" Mobil=Phone, Nettbrett=Tablet, Annet=Other

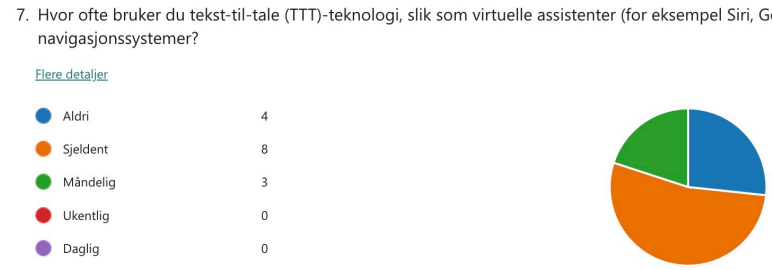

Figure 57: Question 7: "How often do you use text-to-speech (TTS)-technology, such as virtual assistants (e.g. Siri, Google Assistant) or GPS-navigation systems?" Aldri=Never, Sjeldent=Rarely, Månedlig=Monthly, Ukentlig=Weekly, Daglig=Daily

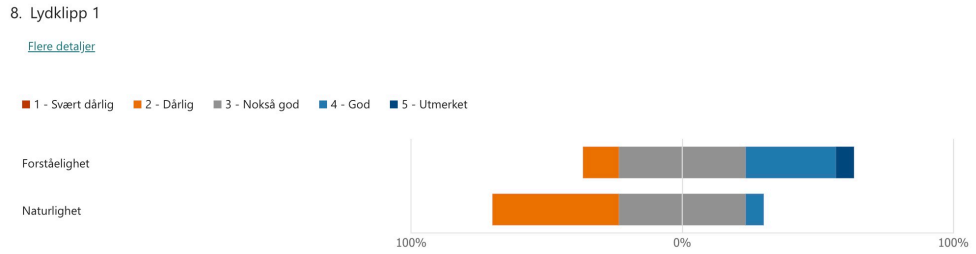

Figure 58: Question 8: "Audio clip 1" Forståelighet=Intelligibility, Naturlighet=Naturalness, Svært dårlig=Bad, Dårlig=Poor, Nokså god=Fair, God=Good, Utmerket=Excellent

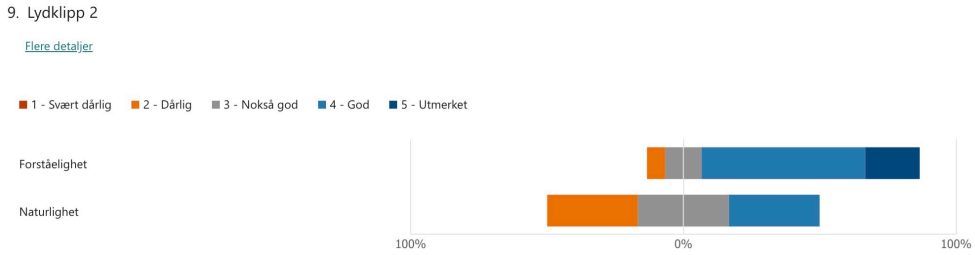

Figure 59: Question 9: ”Audio clip 2”

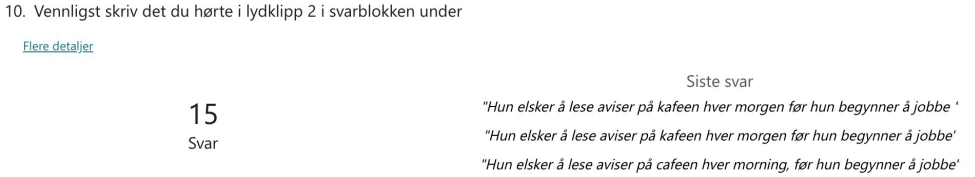

Figure 60: Question 10: ”Please write what you heard in audio clip 2 in the text block below”

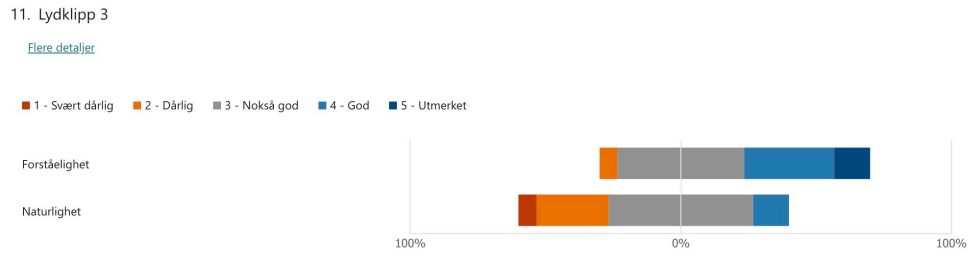

Figure 61: Question 11: ”Audio clip 3”

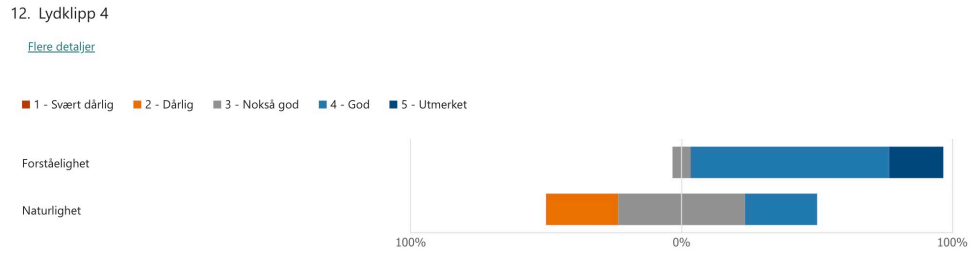

Figure 62: Question 12: ”Audio clip 4”

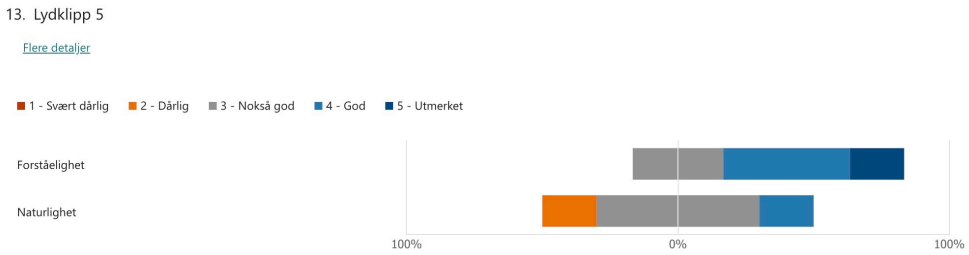

Figure 63: Question 13: ”Audio clip 5”

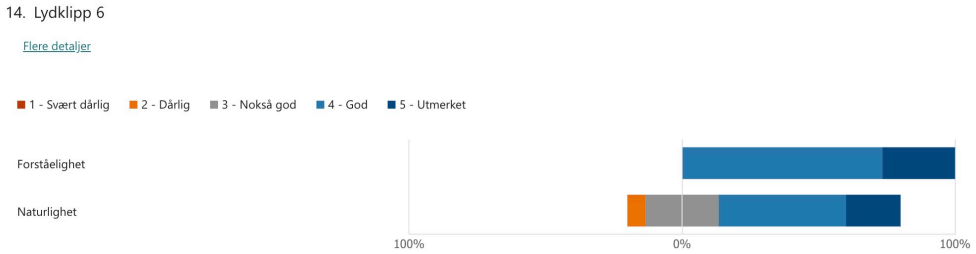

Figure 64: Question 14: ”Audio clip 6”

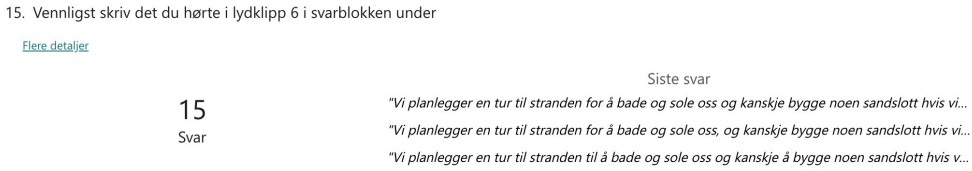

Figure 65: Question 15: ”Please write what you heard in audio clip 6 in the text block below”

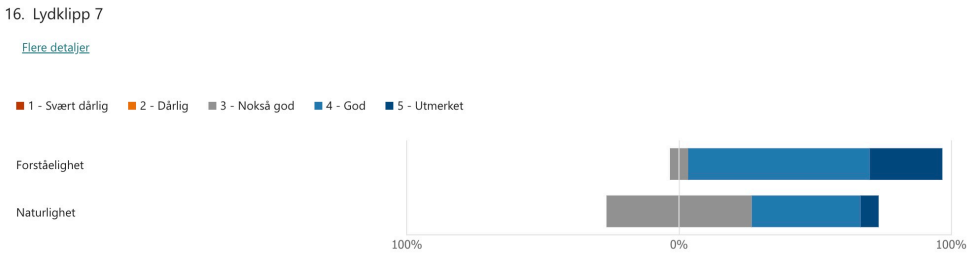

Figure 66: Question 16: ”Audio clip 7”

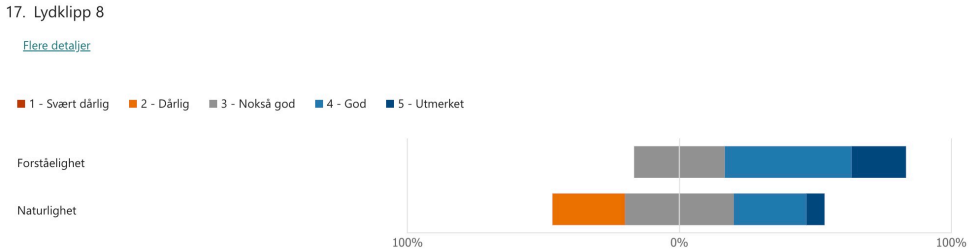

Figure 67: Question 17: ”Audio clip 8”

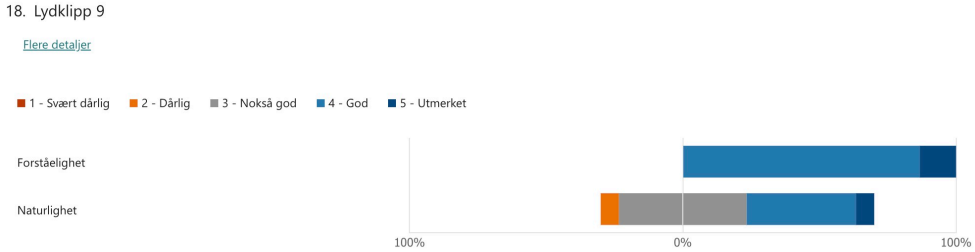

Figure 68: Question 18: ”Audio clip 9”

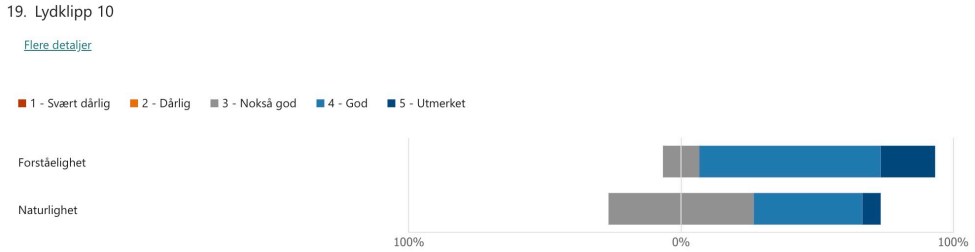

Figure 69: Question 19: ”Audio clip 10”

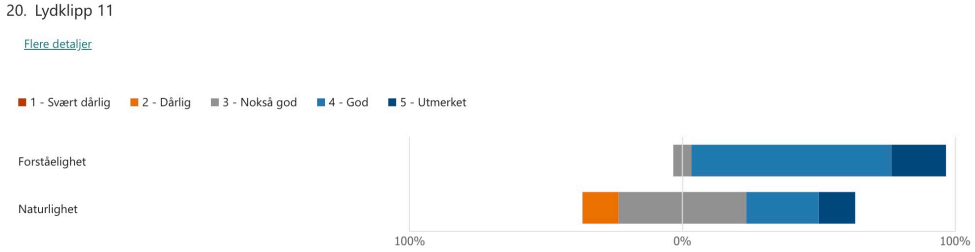

Figure 70: Question 20: ”Audio clip 11”

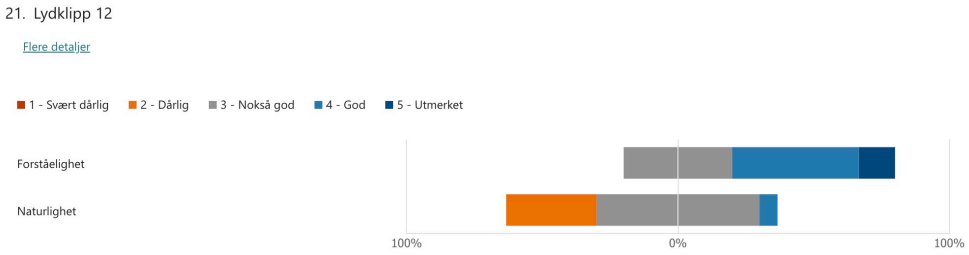

Figure 71: Question 21: ”Audio clip 12”

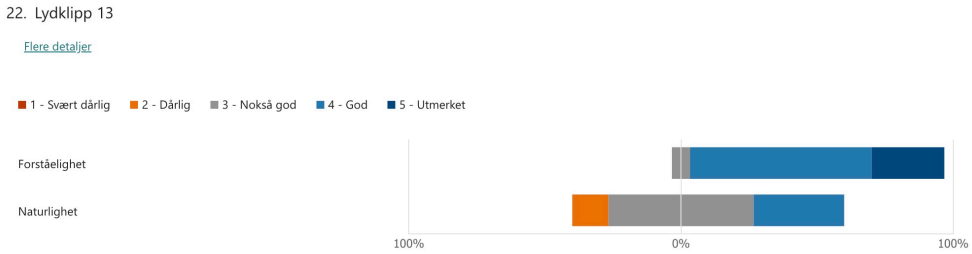

Figure 72: Question 22: ”Audio clip 13”

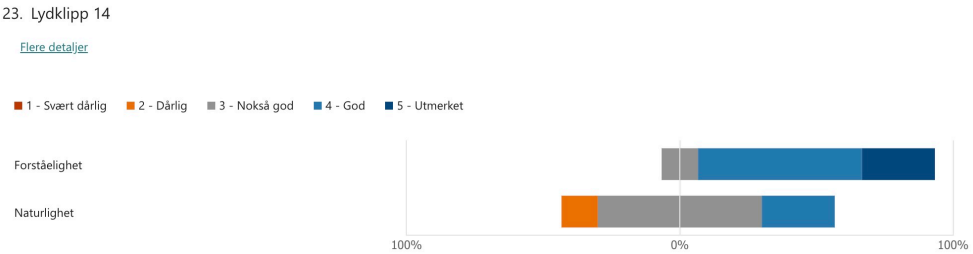

Figure 73: Question 23: ”Audio clip 14”

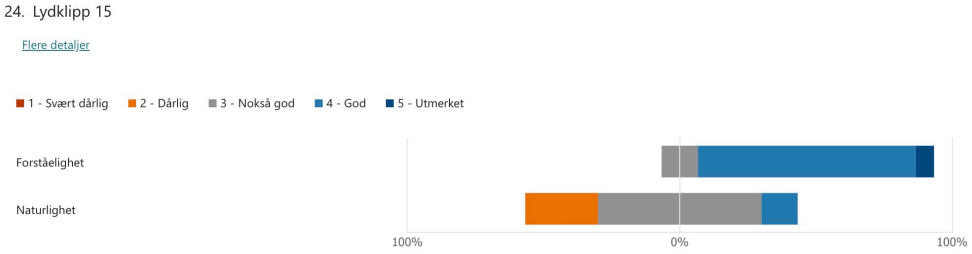

Figure 74: Question 24: ”Audio clip 15”

25. Kryss av på én eller flere påstander som stemmer overens med din generelle forståelse av setningene. Dersom du ikke kjenner deg igjen i noen av påstandene, skriv gjerne din oppfattelse i tekstblokken nederst.

Flere detaljer

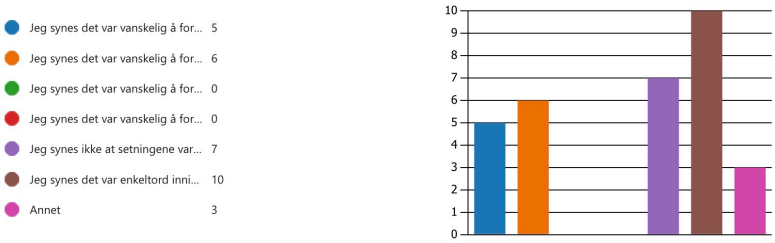

Figure 75: Question 25: ”Select one or more statements that match your general understanding of the sentences. If none of the statements apply to you, please write your interpretation in the text block below.”

TRITA-EECS-EX-2024:571
Stockholm, Sweden 2024