

# Evaluation of wav2vec 2.0 Speech Recognition for the Elderly Frisian Population

**Golshid Shekoufandeh**

A thesis submitted in fulfillment of the requirements of  
*Voice Technology* master's degree

**Supervisor**

Prof. Dr. MATT COLER      Rijksuniversiteit Groningen, The Netherlands

**External Supervisor**

Dr. LYSBETH JONGBLOED      Provincie Fryslân, The Netherlands

**Second Reader**

Dr. VASS VERKHODANOVA      Rijksuniversiteit Groningen, The Netherlands



# Acknowledgments

---

I would like to express my heartfelt gratitude to all those who have contributed to the successful completion of this thesis.

First and foremost, my sincere appreciation goes to Dr. Matt Coler, my supervisor and program director. His exceptional patience, invaluable feedback, and unwavering support have guided me throughout this journey, instilling in me the confidence to undertake this endeavor.

I am deeply thankful to Dr. Lysbeth Jongbloed, my external supervisor, for her guidance and unwavering belief in this project since its inception. Her expertise has played a pivotal role in shaping its outcome.

I am also grateful to my lecturers and teachers for their invaluable guidance, mentorship, and belief in my abilities. Their dedication and passion for teaching have had a profound impact on my academic journey. I am truly thankful for their contributions and the stimulating learning environment they have created.

Furthermore, I want to acknowledge my classmates and cohort members, with special recognition to Dragoş Alexandru Bălan. His invaluable assistance, feedback, and constant encouragement have laid a strong foundation for this project. Without him, this journey would not have been the same.

I thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high-performance computing cluster.

On a personal note, I am immensely grateful to my beloved aunt, Dr. Sepideh Yousefzadeh, for her unwavering support. Her faith in me has been a constant source of inspiration and motivation.

Lastly, I want to express my deepest admiration for my parents, Hengameh and Kamran, for their unwavering love, support, and encouragement throughout my academic journey and the completion of this thesis. From the earliest days of my education, they instilled in me a passion for learning and created a nurturing environment where I could thrive. Their belief in my abilities, the sacrifices they made for my opportunities, and their constant presence have been invaluable sources of motivation, empowering me to overcome challenges. I am truly fortunate to have parents who have played a pivotal role in shaping my educational path, offering unwavering guidance, wisdom, and invaluable advice that have significantly contributed to both my personal and academic growth. I am deeply indebted to them for the profound impact they have had on my life.



# Abstract

---

Automatic Speech Recognition (ASR) converts speech into text. It has become crucial in daily life, as evident through the utility of virtual assistants like Alexa and Siri and other tools that help people. Most publicly available ASR models are designed for the English language. Only a few support Frisian and under-resourced Germanic language. Moreover, none of these models are tailored explicitly for elderly speakers. The lack of adequate ASR resources for the Frisian language poses an intersectional disadvantage for elderly speakers, resulting in significant challenges in developing technologies to address the needs of this community. To address this gap, increasing the availability of training data is necessary. In this study, I propose using data augmentation techniques to augment elderly audio recordings. These augmented datasets will be used to train the wav2vec 2.0 XLS-R model, which has shown promise in Frisian ASR. My co-developed model, fine-tuned from the Facebook XLS-R Wav2Vec2 model, achieved a word error rate (WER) of 15.35% when trained on the Common Voice dataset. The main objective of this research is to investigate the effect of fine-tuning the model using augmented elderly speech data tailored explicitly for Frisian elderly speakers. By integrating this dataset, I expanded the collection of recorded speeches from elderly Frisian individuals, leading to a remarkable 20% improvement in relative WER for Frisian elderly ASR. This study makes a valuable contribution towards tackling the technological hurdles encountered by the local Frisian community. Furthermore, it emphasizes the significance of advancing ASR technologies for languages with limited resources and specific demographic groups. Apart from addressing the research objectives, this study offers essential contextual information, underscores the study's importance, and recognizes the broader implications for ASR research in low-resource languages and elderly ASR.

**Index Terms:** Frisian, elderly speech, speech recognition, self-supervised learning, wav2vec 2.0, XLSR-53, under-resourced language, low-resourced data, data augmentation



# Contents

---

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Statement . . . . .	2
1.2	Motivation . . . . .	3
<b>2</b>	<b>Literature Review</b>	<b>5</b>
2.1	Speech Recognition . . . . .	6
2.1.1	Existing Frisian ASR . . . . .	6
2.1.2	Wav2vec 2.0, a framework for self-supervised learning . . . . .	7
2.1.3	Unsupervised And Self-supervised Cross-lingual Speech Representation Learning . . . . .	8
2.2	Low-Resource Automatic Speech Recognition . . . . .	9
2.3	Voice Conversion Data Augmentation . . . . .	9
2.3.1	Data Augmentation . . . . .	9
2.3.2	Voice Conversion . . . . .	11
2.3.3	StarGANv2-VC . . . . .	12
2.4	Conclusion . . . . .	13
<b>3</b>	<b>Research Question and Hypothesis</b>	<b>19</b>
<b>4</b>	<b>Methodology</b>	<b>21</b>
4.1	Datasets and Prepration . . . . .	22
4.2	Model Training - Original Data . . . . .	23
4.2.1	The architecture of XLS-R model . . . . .	23
4.3	Data Augmentation . . . . .	25
4.4	Model Training - Augmented Data . . . . .	26
4.5	Evaluation Metrics in ASR . . . . .	27
4.6	Ethical issues . . . . .	27
<b>5</b>	<b>Results and Discussion</b>	<b>31</b>
5.1	Results . . . . .	32
5.1.1	Experiment 1 . . . . .	32
5.1.2	Experiment 2 . . . . .	33
5.2	Discussion . . . . .	33
5.2.1	Reflection . . . . .	34
5.2.2	Limitations . . . . .	34

---

<b>6 Conclusion</b>	<b>37</b>
6.1 Culmination . . . . .	38
6.2 Future Work . . . . .	38
6.3 Impact and relevance . . . . .	39
<b>Appendices</b>	<b>41</b>
<b>A Research Proposal</b>	<b>43</b>
<b>B Prompts</b>	<b>55</b>
<b>C Ethical Approval Application</b>	<b>59</b>



## List of Figures

---

4.1	Self-supervised cross-lingual representation learning taken from Conneau et al. (2020). . . . .	24
-----	---	----



## List of Tables

---

2.1	Bibliography of Literature Review References. This table presents a bibliography of references included in the literature review. It is organized into three columns: "Reference" lists the authors and publication year of each reference, "Summary" provides a brief summary of each paper, and "Section" indicates the corresponding section where the paper is mentioned. . . . .	13
4.1	Data Availability for Elderly Population (in minutes) in Common Voice Version 13.0 (Frisian) . . . . .	22
4.2	Data availability for the elderly population in Common Voice Version 13.0 (Frisian) using various techniques. The augmentation includes pitch shifting (p), adding Gaussian noise (gn), applying a band stop filter (b), polarity inversion (pi), and gain augmentation (g). The number following each technique (20 in this case) indicates the percentage of the initial data that has been augmented using that specific method. Additionally, the clean dataset is used alongside the augmented data in all experiments. . . . .	26
5.1	Results for the wav2vec 2.0 XLS-R models fine-tuned on 27.5 minutes of elderly data. The model fine-tuned on greenw0lf/wav2vec2-large-xls-r-1b-frisian serves as the baseline for this project. . . . .	32
5.2	Results obtained from fine-tuning the wav2vec 2.0 XLS-R models after augmenting the data. The names used in the table correspond to the explanations provided in reference 4.2. . . . .	33



CHAPTER 1 

Introduction

---

The world is becoming increasingly digitalized, and technology has impacted our communication. The elderly population is no exception to this trend, as technology has affected their communication habits just as much as any other age group. Much work remains to be done to provide support for small languages. Currently, the focus of automatic speech recognition (ASR) predominantly revolves around larger languages, leaving languages with limited resources behind. However, recent advancements have opened doors for the development of ASR systems tailored to languages with fewer resources. The implications of this advancement are vast, encompassing not only the broader landscape of the Frisian language but also specifically addressing the unique speech patterns of elderly Frisian individuals.

Frysk (West Frisian), a West Germanic language, is predominantly spoken in the province of Fryslân, primarily by individuals of Frisian heritage. Fryslân has a population of 643,000. It is estimated that 74% of the population is capable of speaking Frisian, indicating a total of 400,000 Frisian speakers. Slightly over half of the population speaks Frisian at home. Surveys reveal that approximately 94% of the population can comprehend Frisian, 65% can read it, and only 17% can write in the language (Gorter, 2006).

The linguistic situation makes it expensive and challenging to rely on labeled data to implement ASR systems since they require large amounts of transcribed speech to achieve high performance. To handle this challenge, using semi-supervised models, like wav2vec 2.0, is an appropriate option. Semi-supervised models could be trained with just a small amount of labeled data (as little as one hour), while the rest is unlabeled. Therefore I have chosen to use the wav2vec 2.0 XLS-R model, fine-tuned on Frisian. The existing model has achieved a 15.35% WER (Balan & Shekoufandeh, 2023), representing the baseline for this research.

My focus on elderly speech stems not only from getting the compelling scientific challenge but also from the important social impact. Most state-of-the-art ASR systems exhibit bias toward individuals whose speech patterns differ from those of norm speakers. These norm speakers are usually adults who are highly educated and speak a standardized language variety as their first language (Zhang et al., 2022). Consequently, most research has focused on developing systems catering to non-elderly, non-child speakers, making elderly or children ASR a low-resourced task. This is a concern, especially for the elderly, given that 22% of the population in Fryslân is aged above 65, indicating that approximately one-quarter of the population may not have easy access to ASR systems that accurately recognize their speech. This, in turn, prevents this population from engaging with smart speakers, many IoT devices, and other such innovations. Over time, this could further delegitimize Frisian and hasten replacement with Dutch while for now, it results in a technological gap between those who can use speech technology and those who cannot.

## 1.1. Problem Statement

---

The main objective of this study is to decrease the Word Error Rate (WER) in Frisian elderly ASR systems. Considering that approximately one-fourth of the population in Fryslân consists of elderly individuals who may face challenges in accessing and

---

using technology, the problem is formulated as a speech recognition issue. The proposed approach involves training two models: one using the original data without any augmentations, and another using augmented data. I will compare the performance of these models to assess their effectiveness. To achieve the objective, the following steps will be taken:

1. **Datasets and Preparation:** I select a publicly available dataset and preprocess it.
2. **Model Training - Original Data:** I train the first model on clean data (no augmentation).
3. **Data Augmentation:** I implement various augmentation methods to the dataset, creating augmented data.
4. **Model Training - Augmented Data:** I train the second model on augmented data. Methodologies and timelines are shown in Section 4 and the ethical considerations are discussed.
5. **Evaluation:** Both ASR models will be evaluated using appropriate metrics, with a focus on word error rate (WER).
6. **Analysis and Conclusion:** I analyze the results obtained from the evaluation. The influences of this research are shown in Section 6.

I will investigate augmentation techniques to benefit elderly Frisian speech. Data augmentation is a technique that allows us to increase the amount of data by artificially forming new and different data using existing data.

## 1.2. Motivation

---

The development of an ASR system tailored for the elderly population in a lower-resourced language is driven by the urgent need to address the unique communication challenges faced by this demographic. With advancing age, individuals often encounter changes in their speech patterns and vocal capabilities, posing difficulties in effectively interacting with voice-based technology.

Additionally, in lower-resourced languages like Frisian, the combination of limited resources and linguistic variations further compounds the communication barriers experienced by the elderly population.

The main objective of this paper is to bridge the gaps in automatic subtitling and ASR for the Frisian language. This is achieved by leveraging the wav2vec 2.0 framework, fine-tuning the XLS-R model, and exploring data augmentation techniques. These efforts not only contribute to the field of low-resource ASR but also promote inclusivity for under-resourced languages like Frisian.

The ultimate goal of developing an ASR system specifically designed for the elderly population in a lower-resourced language is to empower and enhance their digital inclusion. By enabling them to fully participate in the increasingly voice-driven technological landscape, this research strives to improve their overall engagement and quality of life.





CHAPTER 2

Literature Review

---

In order to gain a comprehensive understanding of the current landscape of ASR systems developed for low-resourced languages, an extensive literature review was conducted. This chapter presents a thorough summary of the pertinent literature encountered during the research process. The review focused on employing a combination of boolean operators and quoted searches conducted on *Google Scholar*, utilizing relevant keywords to gather a diverse range of sources.

The keywords used in the literature review can be categorized into four main categories:

1. **Speech Recognition:** This category includes keywords such as "speech recognition," "wav2vec 2.0," "xslr-53," "xls-r," "self-supervised learning," and "cross-lingual." These keywords are relevant to the specific techniques and models used in speech recognition research.
2. **Low-resource:** Keywords related to low-resource automatic speech recognition, which is a significant aspect of the study, fall under this category.
3. **Elderly Speech:** This category encompasses keywords such as "elderly speech" and "bias mitigation." These keywords are important for exploring the specific challenges and considerations related to elderly speech in ASR systems.
4. **Voice Conversion and Data Augmentation:** This category includes keywords related to data augmentation techniques, such as "data augmentation," "voice conversion," and "StarGANv2-VC." These techniques are explored in the literature as potential methods to enhance the performance of ASR models.

In this section, I will synthesize the literature I encountered during the review process. The subsequent sections will provide an analysis of the wav2vec 2.0 model, which serves as the baseline for the study (Section 2.1.2). Furthermore, I will explore strategies for improving ASR in low-resourced languages (Section 2.2), discuss various data augmentation techniques that have the potential to enhance model performance (Section 2.3), and finally, provide a summary of the findings from the literature review in Section 2.4.

## 2.1. Speech Recognition

---

In this section, I will examine the current Frisian state-of-the-art discussed in Subsection 2.1.1. Additionally, I will provide an explanation of Wav2vec 2.0 in Subsection 2.1.2. Furthermore, Subsection 2.1.3 will delve into XSLR-53 and XLS-R.

### 2.1.1 Existing Frisian ASR

The report by Robinson-Jones & Scarse (2022) highlights the existing gaps in automatic subtitling and ASR for the Frisian language. Much of the previous work on Frisian is based on the FAME! dataset, which is a bilingual corpus consisting of approximately 18 hours, 33 minutes, and 57 seconds, from which only 14 hours are segments containing speech. This data is split into train, test, and development sub-

sets. In the training set, there are 8.5 hours in Frisian and 3 hours in Dutch (Yilmaz et al., 2016).

Yilmaz et al. (2016) developed an automatic ASR model for Frisian using the FAME! dataset. This model focused on code-switching and achieved a WER of 38.8%. Notably, this publication represents the only publicly available work on Frisian language models to date. WER is an evaluation metric used in ASR, which measures the ratio of errors in a transcript to the total words spoken. Further details on this metric will be provided in Section 4.5 of this paper.

In their work, San et al. (2023) performed fine-tuning on the XLS-R model using Besemah, Nasal, Gronings, and Frisian. For Frisian ASR, they utilized 10 minutes of transcribed data from the FAME! dataset and achieved a reduction in WER from 53.1% to 43%.

The framework of my experiment is based on the greenw0lf/wav2vec2-large-xlsr-1b-frisian model, which is available in the Hugging Face repository. This model has been fine-tuned on the wav2vec 2.0 model and co-developed by me. It achieved a word error rate of 15.35%. I have chosen this model as the baseline for my paper because it is currently the only model that has been fine-tuned specifically on Frisian using the Common Voice version 13 dataset.

### 2.1.2 Wav2vec 2.0, a framework for self-supervised learning

The objective of this study is to investigate the most recent literature on wav2vec 2.0, which is a framework proposed by Baevski et al. (2020) that facilitates self-supervised learning of speech representations. Unlike conventional methods that extract features, this framework masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations.

The model comprises three modules: a multi-layered convolutional feature encoder, a Transformer, and a Quantization module.

The feature encoder, denoted as  $f : \mathcal{X} \rightarrow \mathcal{Z}$ , takes raw audio waveforms as input  $\mathcal{X}$  and produces latent speech representations  $z_1, z_2, \dots, z_{\mathcal{T}}$  for  $\mathcal{T}$  time-steps. The encoder is composed of multiple blocks, each consisting of a temporal convolution, followed by layer normalization and a GELU activation function.

After the feature encoder produces its output, it is fed into a context network that employs the Transformer architecture. The latent representations are then fed into the Transformer, represented as  $g : \mathcal{Z} \mapsto \mathcal{C}$ , which generates representations  $c_1, c_2, \dots, c_{\mathcal{T}}$  that capture information from the entire sequence. Instead of utilizing fixed positional embeddings, which encode absolute positional information, a convolutional layer is adopted that functions as a relative positional embedding. The output of the convolution is added to the inputs, followed by a GELU activation function, and then layer normalization is applied.

The Quantization module, denoted as  $\mathcal{Z} \mapsto \mathcal{Q}$ , utilizes a self-supervised objective to represent the targets by mapping the output of the feature encoder to  $q_t$ . To achieve this, Product quantization is employed, which involves selecting quantized representations from multiple codebooks and concatenating them. Assuming  $\mathcal{G}$  codebooks exist, each with  $\mathcal{V}$  entries, one entry from each codebook is selected and concatenates the resulting vectors  $e_1, e_2, \dots, e_{\mathcal{G}}$ . A linear transformation is applied to this concatenated

vector.

To choose discrete codebook entries in a fully differentiable manner, the Gumbel softmax technique is used.  $\mathcal{G}$  hard Gumbel softmax operations are set up and the straight-through estimator is utilized.

In Yi et al. (2020), the authors used a pre-trained model for solving their ASR task, and they concluded that wav2vec 2.0 had learned basic acoustic units that can be used to compose various languages. They explore the application of the pre-trained wav2vec2.0 model to address the challenges of low-resource Automatic Speech Recognition (ASR) tasks. They investigate the performance of the encoder component of the model across various languages and find that it performs well in ASR tasks. By employing self-supervised training, the authors demonstrate the ability to leverage audio data not only from the target language but also from other languages. They propose that wav2vec2.0 has acquired knowledge of basic acoustic units that can be used to compose diverse languages. Through their analysis, they observe that wav2vec2.0 can dynamically merge fine-grained representations into coarser-grained representations to adapt to the specific target task.

### 2.1.3 Unsupervised And Self-supervised Cross-lingual Speech Representation Learning

The XLSR-53 and XLS-R models, which are both based on the wav2vec 2.0 architecture, have been pre-trained on data from multiple languages, as documented by Conneau et al. (2020); Babu et al. (2021).

XLSR-53, which is unsupervised and has around 300M parameters, was trained on a vast corpus of public training data consisting of approximately 56K hours in 53 languages, as stated by Conneau et al. (2020).

In contrast, the XLS-R model is purpose-built for cross-lingual speech representation and has surpassed previous models in several aspects, including the number of covered languages, volume of training data, and model size. Specifically, the XLS-R model proposed by Babu et al. (2021) is a wav2vec 2.0 model that has been pre-trained on a massive corpus of speech audio from 128 languages worldwide, amounting to almost half a million hours of data. As a result, this pre-trained model serves as the foundation for my model.

The Common Voice corpus, introduced by Ardila et al. (2019), is one of the datasets used to train and evaluate the wav2vec framework. It is a vast collection of over 38 languages with the participation of more than 50,000 individuals. Its online accessibility makes it the most extensive public-domain audio corpus available for ASR, in terms of both language diversity and hour amount. Each language included in the corpus can be downloaded separately as a compressed directory from the Mozilla Common Voice website.<sup>1</sup>

The dataset has been divided in a way that the recordings of each speaker are only present in one of the data splits, ensuring a fair evaluation of speaker generalization. However, this also means that some training sets have very few speakers, making the task even more challenging. The split per language has been made as close as possible

<sup>1</sup><https://commonvoice.mozilla.org/en/datasets>

to 80% for training, 10% for development, and 10% for testing.

---

## 2.2. Low-Resource Automatic Speech Recognition

---

Low-Resource ASR has become a critical area of research due to the limited resources available for developing ASR systems for many languages and dialects spoken by large populations worldwide. By leveraging LSR, accessibility to technology is increased and greater inclusivity for these under-resourced languages and dialects is promoted.

Zheng et al. (2021) focuses on cooperative acoustic and linguistic representation learning for low-resource speech recognition. Their approach leverages the powerful wav2vec 2.0 and BERT models. The authors propose the use of a Representation Aggregation Module and an Embedding Attention Module. These modules facilitate the cooperation between the two pre-trained models, resulting in improved representation learning. The experimental results demonstrate the effectiveness of the proposed Wav-BERT in enhancing low-resource ASR performances across different languages. The authors express their intention to explore more effective modules for incorporating additional types of knowledge and to extend their framework to other pre-trained models, thereby contributing to the advancement of low-resource speech tasks.

In Coto-Solano et al. (2022), three ASR systems were trained for Cook Islands Māori (CIM), an indigenous low-resourced language. One is statistical, based on the Kaldi toolkit, and the other two were based on Deep Learning, DeepSpeech, and XLSR-wav2vec 2.0. Similarly, Phatthiyaphaibun et al. (2022) fine-tuned a pre-trained XLSR-Wav2Vec2 model to train a Thai Automatic ASR system using a newer version of Common Voice corpus.

During the Language Technology for Equality, Diversity, and Inclusion workshop, Srinivasan et al. (2022); Suhasini & Bharathi (2022); Bharathi et al. (2022) fine-tuned models using pre-trained wav2vec 2.0 to improve ASR for Vulnerable Individuals in Tamil, including elderly males, females, and transgender individuals. One key observation they made was that because the pre-trained model used for the system was fine-tuned with the common voice dataset, the model can be trained with one's dataset and used for testing, which can enhance performance.

---

## 2.3. Voice Conversion Data Augmentation

---

In this section, I will explore techniques for enhancing data in scenarios with limited resources. Subsection 2.3.1 will highlight various data augmentation methods. Subsection 2.3.2 will provide an explanation of voice conversion and how it can be employed to generate additional data. Lastly, Subsection 2.3.3 will introduce a voice conversion framework.

### 2.3.1 Data Augmentation

The study conducted by Ragni et al. (2014) delves into the exploration of different schemes aimed at enhancing speech recognition and keyword searching capabilities

for low-resource languages. These schemes include semi-supervised training, acoustic data perturbation, speech synthesis, and multi-lingual processing. The research investigates the utilization of augmented data within tandem and hybrid architectures. Specifically, the tandem architecture was employed for Assamese and Zulu, two low-resource languages, employing both semi-supervised training and acoustic data perturbation individually as well as in combination. The findings demonstrate noteworthy improvements in speech recognition performance for both schemes, with the combined approach yielding the most significant gains, particularly for Zulu.

It has been proven that data augmentation could potentially improve the performance of ASR systems. Sriram et al. (2022) demonstrated that data augmentation can be utilized with Wav2Vec 2.0. They suggested that implementing their proposed model for cross-lingual representation learning in a similar fashion could considerably benefit languages with restricted speech data. In light of this, it is worth mentioning the augmentation methods that can significantly improve ASR for elderly individuals.

Thai et al. (2019) proposed a multistage deep learning approach for low-resource ASR that employs both transfer learning and data augmentation via speech signal distortion and voice conversion. Their study showcased a noteworthy decrease in WER by employing a multi-stage approach, surpassing the results obtained solely by using deep recurrent methods for transfer learning from a high-resource language acoustic model. The approach encompasses three key phases: commencing with initial training through transfer learning from a pre-existing high-resource language acoustic model, followed by weight refinement utilizing a heavily concentrated synthetic dataset, and ultimately fine-tuning the model using a limited synthetic dataset tailored to the target language. As a result, they achieved an impressive 15% reduction in WER compared to the use of deep recurrent methods for transfer learning alone.

Similarly, Jin et al. (2022) utilized SVD-based speech spectrum decomposition to derive spectral and temporal subspace representations because the adversarial data augmentation method they introduced necessitates the use of parallel control and recordings of identical spoken content. In their study, the authors introduced GAN-based data augmentation techniques that consistently outperformed the baseline speed perturbation method. The proposed approaches achieved WER reductions of up to 0.91% absolute (9.61% relative) and 3.0% absolute (6.4% relative) on their datasets.

In their repository, Anidjara et al. (2023) employed basic data augmentation techniques to enhance the performance of the wav2vec 2.0 XLSR-53 model in under-represented languages such as Arabic, Russian, and Portuguese. They used Pitch shift, gaussian noise and band-stop augmentation which is mentioned in the list below as well.

Anidjara et al. (2023), merged their clean data, with 60% randomly chosen each data from these augmentation methods (20% each), and managed to improve the wav2vec 2.0 model by around 30% WER and around 50% Character Error Rate (CER). CER is an ASR evaluation metric that indicates the percentage of characters that were incorrectly predicted. This will be explained further in 4.5.

During my exploration of additional techniques for data augmentation specifically designed for the elderly population, I came across a study by Kaur et al. (2022) Although their work is only tangentially relevant to my own research, they employed

audio recordings in a classification task aimed at detecting instances of falls within bathroom environments, which are particularly susceptible to such events among the elderly. The audio recordings comprised two categories: fall and no-fall. In their approach, they applied several augmentation methods previously employed by Jordal (2023) solely to the no-fall class. This selective application was motivated by the potential risk of compromising the authenticity of the original data through certain transformations, such as completely removing the falling event from the recordings. The authors achieved an accuracy of 86% by developing a solution using a Transformer architecture, which operates on noisy sound inputs derived from bathroom environments and performs classification into the fall or no-fall categories. The authors aside from the augmentations mentioned in the last paragraph implemented some more, and I will also implement some of them in this study.

Some of these data augmentation methods can be seen in the list below, the first two augmentation methods are a result of collective research efforts in the field of machine learning and audio processing.

- **Polarity Inversion:** This augmentation method involves phase inversion, which effectively multiplies the signal by -1, resulting in the reversal of its phase.
- **Gain:** One technique that can enhance a model's ability to handle variations in the overall gain of input audio is amplitude scaling as a form of data augmentation. This involves multiplying the audio signal by a randomly selected amplitude factor, resulting in either a reduction or an increase in volume. By applying this augmentation, the model can become somewhat invariant to changes in audio volume.
- **Pitch Shift:** This augmentation method, was originally proposed by Gfeller et al. (2020). This method involves adjusting the tempo of the audio recording by a certain semitone, with the aim of making the model more resistant to variations in the input data.
- **Gaussian-Noise:** This method was introduced by El Helou & Süssstrunk (2020). This technique aims to improve the model's ability to handle real-world scenarios where the input data may be noisy or contain errors, by enhancing its robustness against such variations.
- **Band-Stop:** This augmentation method presented by Roonizi & Jutten (2021). This technique involves filtering out a specific frequency range from an audio signal using a band-stop filter, which attenuates frequencies within the range while allowing frequencies outside it to pass through.

### 2.3.2 Voice Conversion

To enhance the non-linguistic or para-linguistic aspects of speech, Voice Conversion (VC) is commonly employed. In situations where there is insufficient data to train a strong machine-learning model, low-resource data can be utilized to generate more data that is similar but with minor variations. This technique can help avoid overfit-

ting, where the model becomes too adapted to the limited data available and struggles to generalize to new, unseen data.

In their study Shah Nawazuddin et al. (2020), the authors employed VC as a method of data augmentation to enhance the recognition of children’s speech. They successfully reduced the WER by incorporating VC-based out-of-domain data augmentation. The researchers developed three distinct ASR systems, each trained with different sets of audio recordings. One system utilized children’s audio recordings as the target, another employed adult audio recordings as the target, and the third system combined both datasets. However, due to the scarcity of available data and differences in acoustic characteristics, these factors presented significant challenges. Specifically, when the training solely consisted of children’s speech, the system struggled to accurately recognize adult audio recordings, resulting in a high WER. Conversely, even when adult audio recordings were included in the training, the system’s performance in recognizing children’s speech still experienced noticeable degradation compared to the dedicated children’s system. On a positive note, the third ASR system achieved satisfactory WERs for both test sets. This outcome emphasizes the effectiveness of augmenting the training data with out-of-domain samples, as it significantly improves the accuracy of ASR for children’s speech.

Further, Kim et al. (2021) concentrated on improving a recognition system that struggles with recognizing outlier voices, such as the elderly. They proposed age-to-age voice translation using linguistic-coupled information to enhance ASR performance for elderly individuals. In addition to other methods, the authors also utilized the conventional vocal tract length normalization (VTLN) technique proposed by Eide & Gish (1996) to assess the improvement in performance. They applied VTLN to normalize the speech data of elderly males and females based on the vocal tracts of adult males and females, respectively. The results showed that VTLN had some effectiveness in improving the recognition performance for elderly male speech, but it exhibited limited improvement for female speech. This indicates that the normalization method has limitations in enhancing the ASR system’s performance for the elderly.

However, the most impressive results were achieved using an age-to-age voice translation approach and merging linguistic identifiers. This method achieved a CER of 19.21% for adult males and 14.35% for adult females, surpassing all other techniques employed in the study.

### 2.3.3 StarGANv2-VC

During my extensive research on voice conversion frameworks, I discovered an impactful framework called StarGAN, proposed by Choi et al. (2018) StarGAN addresses the challenge of multi-domain image-to-image translation within a single dataset. Notably, Choi et al. (2020) expanded upon this work by enabling StarGAN to convert images from one reference to multiple target images, showcasing its versatility.

The advancements made by Choi et al. (2020) in the field of image-to-image translation directly influenced the work of Li et al. (2021) in developing an unsupervised, nonparallel framework for voice conversion. This innovative framework eliminates the need for explicit training and enables any-to-many, cross-lingual voice conversion



capabilities. The achievements of StarGAN and subsequent research have opened up new possibilities and avenues for voice conversion, pushing the boundaries of what is possible in this domain.

## 2.4. Conclusion

In conclusion, the existing research on ASR for the Frisian language reveals significant gaps in this field. Most of the work done on Frisian is based on a bilingual dataset called FAME!, which in the training set includes only 8.5 hours of speech in Frisian. The available models for Frisian, such as the one developed by Yilmaz et al. (2016), achieved a WER of 38.8%. However, recent advancements in ASR, particularly the fine-tuning of the XLS-R model on augmented Frisian data by San et al. (2023), have shown promise in reducing the WER from 53.1% to 43%.

For this paper, I have chosen the greenwolf/wav2vec2-large-xls-r-1b-frisian model, fine-tuned on the wav2vec 2.0 model, as the baseline. This model achieved a WER of 15.35% and is the only publicly available model fine-tuned on Frisian using the version 13.0 Common Voice dataset. The wav2vec 2.0 framework, proposed by Baevski et al. (2020), has shown significant potential for self-supervised learning of speech representations, utilizing a multi-layered convolutional feature encoder, a Transformer, and a Quantization module.

Furthermore, the use of pre-trained models like XLSR-53 and XLS-R, based on the wav2vec 2.0 architecture, has been successful in cross-lingual speech representation learning. These models have been pre-trained on vast corpora of speech audio from multiple languages, enabling better performance and generalization across languages.

Data augmentation techniques, such as polarity inversion, gain, pitch shift, Gaussian noise, and band-stop filtering, have proven effective in improving the performance of ASR systems. Voice conversion has also been used as a method of data augmentation to generate additional data with minor variations, thereby enhancing the model's ability to handle diverse speech characteristics.

Table 2.1, provides a comprehensive compilation of all the references cited in the context of low-resourced languages and vulnerable individuals.

Table 2.1: Bibliography of Literature Review References. This table presents a bibliography of references included in the literature review. It is organized into three columns: "Reference" lists the authors and publication year of each reference, "Summary" provides a brief summary of each paper, and "Section" indicates the corresponding section where the paper is mentioned.

Reference	Summary	Section
Baevski et al. (2020)	The paper introduces <b>Wav2vec 2.0</b> , a self-supervised learning framework for speech representations that mask segments of the raw waveform. It achieves state-of-the-art results on the Librispeech benchmark, outperforming previous methods with <b>limited labeled data</b> and showing potential for scalability and architectural enhancements.	2.1

Conneau et al. (2020)	The paper explores unsupervised cross-lingual speech representations learned from raw waveforms. Pretraining on multilingual data improves performance, especially for <b>low-resource languages</b> . Fine-tuning the model on multiple languages simultaneously enables a competitive multilingual speech recognition model, indicating shared capacity across languages, particularly among related ones.	2.1
Babu et al. (2021)	The paper introduces <b>XLS-R</b> , a large-scale cross-lingual speech representation model trained on a vast amount of multilingual data. XLS-R achieves state-of-the-art results in speech translation, speech recognition, and language identification tasks, surpassing previous approaches and demonstrating its generalization ability. With its capacity and scalability, XLS-R has the potential to enhance speech processing for a wide range of languages and applications.	2.1
Yi et al. (2020)	The paper explores the application of pre-trained wav2vec2.0 models to <b>low-resource</b> speech recognition tasks in different languages. The experiments demonstrate significant improvements in performance compared to previous work, with gains of over 20% in multiple languages, including a remarkable 52.4% improvement in English. Coarse-grained modeling units, such as subwords or characters, yield better results than fine-grained units like phones or letters. The study suggests that wav2vec2.0's self-supervised training enables it to leverage audio data from various languages, and the model dynamically adapts its representation to fit the specific task at hand.	2.1
Ardila et al. (2019)	<b>Common Voice</b> is a crowd-sourced, multilingual speech <b>corpus</b> designed for speech technology research. It is the largest public domain dataset for ASR and has been collected and validated through community effort. The corpus includes data from 38 languages, and experiments using Common Voice have shown significant improvements in ASR performance for various target languages.	2.1
Yilmaz et al. (2016)	The paper presents a new speech database with 18.5 hours of annotated radio broadcasts in <b>Frisian</b> and Dutch. The database covers almost 50 years and captures <b>code-switching</b> between Frisian and Dutch. It includes manual annotations for orthographic transcription, speaker identities, dialect information, code-switching details, and background noise/music.	2.1
Yilmaz et al. (2016)	The paper presents a code-switching ASR system for the Frisian language. The study aims to explore the influence of language switching on modern ASR systems and develop a robust recognizer. A bilingual deep neural network <b>DNN-based ASR system</b> is designed to handle code-switching speech, and the impact of bilingual DNN training is investigated within this context.	2.1
San et al. (2023)	Recent research shows that fine-tuning a pre-trained transformer model for ASR can be effective with just 10 minutes of transcribed speech, given the availability of vast amounts of text data. However, relying solely on a lexicon does not significantly improve ASR performance, while combining lexica and language models from larger text corpora can reduce the WER to 39% on average for certain language pairs. This suggests that fine-tuning with <b>minimal speech data</b> alongside a sufficiently large text corpus holds promise for achieving more accurate transcriptions.	2.1

Zheng et al. (2021)	The paper addresses the challenge of unifying acoustic and linguistic representation learning for low-resource speech recognition. Existing approaches cascade pre-trained models, but fail to address the representation discrepancy and may lead to catastrophic forgetting. The proposed method, Wav-BERT, combines wav2vec 2.0 and BERT into an end-to-end trainable framework, utilizing a Representation Aggregation Module and an Embedding Attention Module to enhance representation learning. Experimental results demonstrate that Wav-BERT outperforms previous methods and achieves state-of-the-art performance in <b>low-resource</b> speech recognition.	2.2
Coto-Solano et al. (2022)	This paper describes the transcription, data preparation, and training of ASR for Cook Islands Maori, an Indigenous language of Polynesia. The best-performing ASR systems achieve a CER of 6 and a WER between 18 and 23 for short utterances, showing promise for accelerating transcription. The study demonstrates that deep learning can be successful in <b>low-resource</b> environments and with minority/Indigenous languages, and further efforts are planned to expand the dataset and integrate ASR into the documentation pipeline.	2.2
Phatthiyaphaibun et al. (2022)	A Thai ASR model was developed by fine-tuning a pre-trained XLSR-53 Wav2Vec 2.0 model with an updated CommonVoice corpus, utilizing a trigram language model to improve performance. This approach resulted in a lower word error rate compared to previous works, showcasing the effectiveness of the language model in enhancing ASR accuracy.	2.2
Srinivasan et al. (2022)	This paper describes an automatic speech recognition model trained on the Facebook <b>XLSR-53</b> and <b>XLS-R</b> Wav2Vec2 models using the <b>Common Voice Dataset</b> . The model achieved a word error rate of 39.4512% and focuses on making speech recognition more accessible to the Tamil-speaking population. The authors suggest further fine-tuning with a larger dataset and diverse accents to improve transcription accuracy.	2.2
Bharathi et al. (2022)	This paper provides an overview of a shared task on automatic speech recognition in the Tamil language. The task involved recognizing and evaluating spontaneous Tamil speech data from <b>elderly</b> and transgender individuals gathered in public locations. <b>Transformer-based models</b> were used by participants, and the paper discusses the results obtained using various pre-trained transformer models.	2.2
Suhasini & Bharathi (2022)	An ASR system is developed to address Tamil conversational speech data from <b>elderly individuals</b> and transgender individuals. The system utilizes a <b>pre-trained model</b> and achieves a word error rate of 39.65%. Future improvements could involve training the model with a customized dataset to further enhance performance in recognizing speech from the elderly and transgender populations.	2.2
Gfeller et al. (2020)	This paper introduces SPICE, a <b>self-supervised pitch estimation algorithm</b> for monophonic audio. The SPICE model, trained without labeled data, demonstrates competitive performance in pitch estimation compared to CREPE, a fully-supervised model. The SPICE model is publicly accessible as a Tensorflow Hub module.	2.3

El Helou & Süssstrunk (2020)	The paper presents a theoretical framework for fusion <b>denoising</b> , an optimal denoising solution integrated into a deep learning architecture. Experimental results demonstrate that the proposed Fusion Net outperforms existing methods on higher unseen noise levels, providing improved real image denoising performance.	2.3
Roonizi & Jutten (2021)	This contribution explores the extension of smoothness priors and quadratic variation regularization to <b>band-stop smoothing filters</b> (BSSFs). The paper demonstrates that the optimization approaches effectively control the cutoff frequencies and the order of the BSSFs, resulting in improved performance and the ability to create BSSFs with sharp transition bands for high-performance applications.	2.3
Sriram et al. (2022)	This work addresses the challenge of applying <b>self-supervised learning</b> to domains with <b>limited available data</b> , particularly in languages with a scarcity of unlabeled data. By leveraging <b>data augmentation</b> for Wav2Vec 2.0 pretraining and introducing improvements to the model, the proposed approach achieves a significant relative WER improvement of up to 13% compared to Wav2Vec 2.0 on the Librispeech test-clean/other datasets.	2.3
Thai et al. (2019)	This paper addresses the challenge of improving ASR accuracy for <b>low-resource languages</b> by investigating various methods of acoustic modeling and <b>data augmentation</b> . The proposed approach utilizes transfer learning, synthetic data generation, and fine-tuning techniques, resulting in a significant 15% reduction in word error rate compared to deep recurrent methods and a 19% improvement over traditional frameworks using deep convolutional approaches.	2.3
Jin et al. (2022)	This paper introduces personalized adversarial <b>data augmentation</b> approaches using speaker-dependent generative adversarial networks (GAN) for dysarthric and <b>elderly</b> speech recognition tasks. The experiments conducted on various datasets demonstrate improved coverage and model generalization compared to conventional methods, such as tempo or speed perturbation and SpecAugment, in both hybrid TDNN and end-to-end Conformer ASR systems.	2.3
Kaur et al. (2022)	This work proposes a Transformer-based deep learning model for fall detection among the elderly using ambient sound input. The approach offers a non-wearable, non-intrusive, and scalable solution, particularly suitable for bathroom environments where other techniques may not be practical. In this paper, the authors used <b>data augmentation</b> to increase the size of the dataset by applying various transformations to the original data.	2.3
Shahnawazuddin et al. (2020)	This work addresses the challenge of limited speech data for developing ASR systems for children. The paper proposes the use of a GAN-based <b>voice conversion</b> to augment adult speech data and make it perceptually similar to children’s speech, resulting in improved recognition rates. Additionally, the study explores the use of cycle-consistent GAN for out-of-domain <b>data augmentation</b> and incorporates speaking-rate adaptation to further enhance children’s speech recognition.	2.3

---

Kim et al. (2021)	This paper addresses the low-performance issue of ASR for the <b>elderly</b> by proposing a neural network-based <b>voice conversion</b> framework. The approach includes unsupervised phonology clustering to extract linguistic information and age-to-age voice translation to enhance speech recognition accuracy. The proposed method, which can be used with any commercial or open ASR system, demonstrates significant improvements in recognizing elderly speech without directly modifying the speech recognizer.	2.3
Choi et al. (2018)	This paper introduces StarGAN, a novel approach for <b>image-to-image translation</b> that can handle multiple domains using a single model. Unlike existing methods, StarGAN allows simultaneous training of multiple datasets with different domains, resulting in superior quality translations and the ability to flexibly translate images to any desired target domain.	2.3
Choi et al. (2020)	This paper introduces StarGAN v2, an <b>image-to-image translation</b> framework that addresses the challenges of diversity and scalability in multiple visual domains. It achieves significantly improved results compared to existing methods, as demonstrated through experiments on CelebA-HQ and a new animal faces dataset. The release of AFHQ provides a valuable resource for assessing image-to-image translation models.	2.3
Li et al. (2021)	This paper presents an unsupervised non-parallel <b>many-to-many voice conversion</b> method using StarGAN v2. The model achieves significant improvements in voice conversion compared to previous models, demonstrating the ability to handle various conversion tasks and produce natural-sounding voices without the need for text labels. Additionally, the model is fully convolutional and can perform real-time voice conversion with a fast vocoder.	2.3



CHAPTER 3 

Research Question and Hypothesis

As it can be seen in Section 2.4 of the literature review, it is shown that data augmentation techniques have been successful in improving the accuracy of low-resource language ASR systems. However, there is a lack of research on the use of data augmentation for elderly Frisian speech. To address this gap, I pose the following research question:

To what extent do data augmentation techniques improve the accuracy and reliability of automatic recognition systems for elderly Frisian speech?

Previous studies have demonstrated the effectiveness of basic data augmentation methods, such as the band-stop filter Roonizi & Jutten (2021), Gaussian noise El Helou & Süssstrunk (2020), and pitch shift Gfeller et al. (2020), in reducing the WER in underrepresented languages Anidjara et al. (2023). These methods have also been tested by Kaur et al. (2022) as data augmentation techniques for classifying elderly data in a task focused on detecting instances of falls in bathroom settings, where they improved the classification accuracy. It is plausible that these methods could have a similar positive impact on speech recognition tasks.

There are several advantages to utilizing basic data augmentations. Firstly, their practicality is noteworthy. The data preparation process requires only a relatively short amount of data, typically measured in minutes, as opposed to the need for recording additional individuals. This saves valuable time, energy, and resources that would otherwise be expended.

In addition to practicality, another advantage is the efficiency of basic data augmentations. These augmentation methods can be applied using minimal computational resources. Consequently, they offer a computationally fast approach to data augmentation, enhancing the overall efficiency of the process.

However, if the use of data augmentation fails to support the hypothesis, it suggests that the chosen data augmentation techniques may not improve ASR for elderly individuals. In such cases, exploring more complex data augmentation methods becomes necessary. For instance, synthetic data augmentation Thai et al. (2019) and personalized adversarial data augmentation Jin et al. (2022) could be considered. Synthetic data augmentation involves generating speech data synthetically based on the unique characteristics of each individual's speech patterns. On the other hand, adversarial data augmentation aims to augment the existing training data by using adversarial techniques. These alternative data augmentation methods have the potential to increase the diversity of training data and enhance the robustness of the ASR system.

It is important to acknowledge that factors other than data augmentation may influence the accuracy of ASR systems for elderly speech, such as age-related changes in speech production or microphone placement. Thus, if the research question is falsified, it presents an opportunity to re-evaluate the factors impacting the accuracy and reliability of ASR systems for elderly Frisian speech and explore alternative strategies for improvement. This may also spark a further investigation into understanding the unique characteristics of elderly Frisian speech and developing tailored solutions to enhance automatic recognition systems for this specific demographic.



CHAPTER 4 

Methodology

---

This chapter outlines the methodology employed to validate the hypothesis. Firstly, the datasets and the necessary data preparation steps are described in Section 4.1. Subsequently, Experiment 1 is detailed in Section 4.2, which focuses on the training process using the unaugmented data. The data augmentation procedure is then presented in Section 4.3. Moving forward, Experiment 2 is explained in Section 4.4, where both augmented and unaugmented data are utilized for training. In order to properly evaluate the ASR systems addressed in this paper, it is crucial to introduce and elaborate on the metrics employed. These metrics are comprehensively discussed in Section 4.5. In Section 4.6 I mention the ethical issues that could be issued in this study.

## 4.1. Datasets and Prepration

I acquired the Common Voice corpus, specifically version 13.0, which contains 67 hours of speech in Frisian. The corpus was obtained through crowdsourcing, where individuals utilized the Common Voice website to record their voices while reading sentences displayed on the screen (Ardila et al., 2019). This approach enabled a large and diverse set of recordings from various speakers that represented the target population, which was valuable for the project’s objectives. The Frisian Common Voice dataset version 13.0 comprises 55.6 minutes of elderly speech. Table 4.1 provides the elderly data available in each subset.

Table 4.1: Data Availability for Elderly Population (in minutes) in Common Voice Version 13.0 (Frisian)

<b>Split</b>	<b>Decades</b>	Female	Male	<b>Total</b>
<b>Train</b>	sixties	18.96m	6.06m	25.02m
	seventies	0m	2.35m	2.35m
	eighties	0	0.08m	0.08m
	<b>Total</b>	18.96m	8.49m	<b>27.45m</b>
<b>Dev</b>	sixties	15.26m	3.44m	18.7m
	seventies	0m	1.77m	1.77m
	eighties	0m	1.09m	1.09m
	<b>Total</b>	15.26m	6.3m	<b>21.56m</b>
<b>Test</b>	sixties	4.1m	2.2m	6.3m
	seventies	0.18m	0.11m	0.29m
	eighties	0m	0m	0m
	<b>Total</b>	4.28m	2.31m	<b>6.59m</b>

The speech data undergo preprocessing procedures to enhance its quality and optimize it for training ASR models. The dataset is obtained from the Common Voice version 13 and consists of multiple '.tsv' files, among which the 'train.tsv', 'eval.tsv', and 'test.tsv' files are of particular interest as they represent the train, development,

and test sets respectively.

Each of these files contains several metadata columns, encompassing the recording ID, audio file path, spoken sentence, and speaker attributes such as age, gender, and accent. To focus on the task of speech recognition irrespective of the speaker, the datasets are filtered to exclusively extract recordings from elderly speakers. Subsequently, all metadata columns, with the exception of the audio path and spoken sentence, are discarded. The remaining columns are then subjected to preprocessing. To adapt to the requirements of this project, where '.csv' files are utilized, quotation marks are introduced around sentences to account for potential commas since they are employed as separators in '.tsv' files.

The text labels associated with the speech samples necessitate simplification to align the speech features obtained from wav2vec 2.0 with characters through the use of a CTC loss function (Graves, 2012). In the text normalization phase, the first step involves the removal of special characters that do not represent specific sounds, such as punctuation marks (.,?!;:). Additionally, all characters are converted to lowercase since uppercase and lowercase letters carry the same meaning during speech decoding. The resulting preprocessed sentences exclusively comprise characters from the Frisian alphabet and whitespace to separate words.

The vocabulary utilized encompasses Frisian characters and whitespace, with the whitespace being transformed into a distinctive token ('|'). Furthermore, two additional tokens are introduced to facilitate CTC decoding: an unknown token representing unidentified characters and a blank token to aid in the decoding of words with consecutive letters.

Subsequently, the audio undergoes preprocessing. The raw audio files, which exhibit varying durations, are read with a sampling rate of 16kHz, which is deemed sufficient for audio processing.

## 4.2. Model Training - Original Data

---

The first ASR model is trained using the original data without any additional augmentations on greenwolf/wav2vec2-large-xls-r-1b-frisian, which is based on facebook/wav2vec2-xls-r-1b, and trained on Frisian using the Common Voice dataset. I did this because this model has already been fine-tuned on Frisian and would get better results than the model not fine-tuned at all on Frisian. This model serves as a baseline for comparison.

### 4.2.1 The architecture of XLS-R model

The Wav2Vec2.0 XLS-R model is an enhanced version of the original Wav2Vec2.0 architecture, designed specifically for the XLS-R framework. Building upon the principles discussed in Subsection 2.1.3 of the literature review, the Wav2Vec2.0 XLS-R model leverages cross-lingual transfer learning from high-resource languages to improve representations for low-resource languages with limited unlabeled data. Notably, the XLSR-53 variant, the largest model in the XLS-R family, was trained on

approximately 50,000 hours of publicly available training data across 53 languages, consisting of around 300 million parameters, as reported by Conneau et al. (2020).

By tailoring the original Wav2Vec2.0 architecture for the XLS-R framework, the Wav2Vec2.0 XLS-R model harnesses the benefits of unsupervised pretraining on audio data. This empowers the model to acquire rich representations of speech by leveraging a large volume of unannotated audio data during the self-supervised pretraining phase. Consequently, the Wav2Vec2.0 XLS-R model demonstrates exceptional performance in speech recognition, transcription, and other speech-related natural language processing tasks.

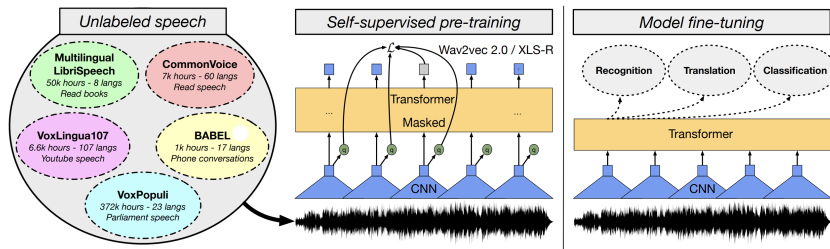


Figure 4.1: Self-supervised cross-lingual representation learning taken from Conneau et al. (2020).

The architecture depicted in Figure 4.1 is now presented and explained. Firstly, the audio data undergoes preprocessing, as described in Section 4.1. The hyperparameters of the model need to be adjusted based on the dataset before the training starts. Although there are multiple hyperparameters that can be tuned, I focused solely on experimenting with the initial learning rate of the AdamW optimizer, which is a variant of Adam that separates weight decay from the gradient update. For the selection of values, I followed the range specified in Section 4.3 of Conneau et al. (2020).

In the pretraining phase, the model is trained on unlabeled speech data. The input to the model is a waveform, which is processed by a Convolutional Neural Network (CNN) serving as a feature encoder. CNN learns hierarchical representations of the input data through convolutional layers. The output of this layer is then passed through a Quantization Module and a Transformer.

The Quantization Module maps continuous speech features to discrete representations using learned codebooks. These codebooks contain prototype vectors representing possible quantized values. The module quantizes the input by finding the closest prototypes in the codebooks.

The outputs of the modules are the Quantized and Contextual representations, which are used to minimize the Loss function. The Loss function consists of two components: Contrastive Loss ( $\mathcal{L}_m$ ) and diversity loss ( $\mathcal{L}_d$ ). The total Loss function is calculated as  $\mathcal{L} = \mathcal{L}_m + \alpha\mathcal{L}_d$ .

The Contrastive Loss ( $\mathcal{L}_m$ ) aims to minimize misidentifications between the true latent representations and a set of K distractors. Distractors refer to other represen-

tations from the same utterance. It is computed using a logarithmic equation.

$$\mathcal{L}_m = -\log \frac{\exp \frac{\text{sim}(c_t, q_t)}{K}}{\sum_{\tilde{q} \sim Q_t} \exp \frac{\text{sim}(c_t, \tilde{q})}{K}} \quad (4.1)$$

The diversity loss ( $\mathcal{L}_d$ ) encourages the use of different codebook representations. It involves a softmax distribution over the codebook entries.

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^G -H(\tilde{p}_g) = \frac{1}{GV} \sum_{g=1}^G \sum_{v=1}^V \tilde{p}_{g,v} \log \tilde{p}_{g,v} \quad (4.2)$$

This training process, incorporating the quantization module, codebooks, and loss function, enables the model to learn meaningful representations from speech data, leading to improved performance.

During the fine-tuning phase within this framework, the data undergoes a similar process. The waveform serves as the input to a CNN, which acts as the feature encoder. The transformed data is then passed through a transformer. The resulting output from the model can be utilized for various tasks, including recognition, translation, and classification.

Through extensive experimentation with various values, I fine-tuned the model using my 27 minutes of data and achieved the best outcome with a learning rate of 3e-5. The training time for 27 minutes of data takes less than 25 minutes.

To conduct the fine-tuning of all XLS-R models, I utilized an Nvidia Tesla A100 Tensor Core GPU with 40GB of VRAM and 512GB of memory.

### 4.3. Data Augmentation

To address the limited availability of training data and potentially improve the robustness of the ASR system, data augmentation techniques are applied. These techniques involve artificially expanding the training data by introducing variations such as background noise, pitch shift, and band rejection. Augmented data is generated by applying these modifications to the original speech samples.

According to Anidjara et al. (2023), Each of these augmentations, namely Pitch-Shift, Band-Stop, and Gaussian-Noise, has its unique approach to altering the input audio data. Pitch-Shift uniformly modifies the semitone of the entire audio recording by a fixed amount, while Band-Stop removes a specific frequency range from the audio signal. On the other hand, Gaussian-Noise alters the amplitude of the entire audio recording slightly, simulating the presence of white noise. Despite their differences, each of these augmentations is crucial in enhancing the model’s accuracy. Additionally, two other data augmentations are tested to see their effect on the model, Gain, and Polarity inversion. Polarity inversion is a technique used to modify audio signals by reversing the polarity of the waveform. In the context of speech processing, the polarity of a waveform determines whether the air pressure increases or decreases over time. By inverting the polarity, the positive and negative amplitudes of the waveform are switched, resulting in a mirrored version of the original signal. Gain refers to the

amplification or attenuation of the audio signal’s amplitude. It is a parameter that determines the volume or loudness of the speech waveform. By adjusting the gain, the overall energy level of the signal can be modified.

I used 20% of the data acquired by each data augmentation method and made sure that I am not using the same audio recording multiple times because this might overfit that model. Alongside the augmented data, I also include 100% of the clean data. Table 4.2, the amount of data that has been generated in each experiment. The name of the experiments indicates what kind of data augmentation has been done on the data in that experiment.

Table 4.2: Data availability for the elderly population in Common Voice Version 13.0 (Frisian) using various techniques. The augmentation includes pitch shifting (p), adding Gaussian noise (gn), applying a band stop filter (b), polarity inversion (pi), and gain augmentation (g). The number following each technique (20 in this case) indicates the percentage of the initial data that has been augmented using that specific method. Additionally, the clean dataset is used alongside the augmented data in all experiments.

<b>Experiments</b>	<b>Decades</b>	female	male	<b>Total</b>
<b>20p20gn20b</b>	sixties	26.76	8.06	34.82
	seventies	0	3.53	3.53
	eighties	0	0.15	0.15
	<b>Total</b>	26.76	11.74	<b>38.5</b>
<b>20p20gn20b20pi</b>	sixties	34.39	10.41	44.8
	seventies	0	4.21	4.21
	eighties	0	0.15	0.15
	<b>Total</b>	34.39	14.77	<b>49.16</b>
<b>20p20gn20b20pi20g</b>	sixties	39.1	12.36	51.46
	seventies	0	3.75	3.75
	eighties	0	0.23	0.23
	<b>Total</b>	39.1	16.34	<b>55.44</b>

#### 4.4. Model Training - Augmented Data

The second ASR model is trained using augmented data. The augmented data that is achieved by implementing the methods mentioned in Section 4.3 is used to fine-tune the greenwolf/wav2vec2-large-xls-r-1b-frisian. Through extensive experimentation with various values, I fine-tune the model in multiple stages and used different learning rates.

To conduct the fine-tuning of all XLS-R models, I utilize an Nvidia Tesla A100 Tensor Core GPU with 40GB of VRAM and 512GB of memory. The training time for

augmented data increases as the number of augmented data increases and for around 1 hour of data that I ended up with, the training time is around 45 minutes.

This model is then evaluated to determine if data augmentation techniques have a positive impact on the ASR performance compared to the baseline model.

## 4.5. Evaluation Metrics in ASR

---

The first metric is Word Error Rate (WER), which serves as a standard measure for assessing accuracy. WER quantifies the disparity between the recognized words produced by the ASR system and the reference transcription, which represents the ground truth. It involves calculating the number of insertions, deletions, and substitutions required to convert the recognized words into the reference transcription. WER is typically expressed as a percentage, reflecting the error rate relative to the total number of words in the reference transcription. A lower WER signifies superior accuracy and improved performance of the ASR system. Researchers and developers commonly utilize WER as a benchmark for comparing different ASR models, techniques, and system variations.

The second metric discussed is Character Error Rate (CER), which focuses on evaluating transcription accuracy at the character level. CER measures the dissimilarity between the recognized output generated by the ASR system and the reference transcription. It involves tallying the number of insertions, deletions, and substitutions needed to transform the recognized characters into the reference transcription. While the operations involved in calculating CER are similar to those used in WER, they operate at the character level rather than the word level. CER is also expressed as a percentage, indicating the proportion of errors in relation to the total number of characters in the reference transcription. Like WER, a lower CER indicates higher accuracy and improved performance in capturing the details of the spoken language. Researchers and developers commonly employ CER as a metric for evaluating different ASR models, techniques, and system variations.

## 4.6. Ethical issues

---

As technology continues to advance, ASR systems offer promising solutions for enhancing communication and accessibility for older individuals who for example, face challenges in verbal communication. However, the integration of ASR systems in this context raises profound ethical questions that warrant careful examination. How do we ensure the privacy and security of personal information shared through these systems? How can we mitigate potential biases and ensure fair treatment in the design and implementation of ASR systems for elderly individuals with limited resources? This Section aims to explore the ethical dimensions surrounding the development and deployment of ASR systems tailored to the needs of elderly individuals with limited resources, addressing the ethical responsibilities, considerations, and potential solutions that can ensure the well-being and empowerment of this vulnerable population.

The study will commence by utilizing the publicly available Common Voice Corpus 13.0 dataset, which was collected and validated through Mozilla's Common Voice initiative. This dataset offers a sizeable collection of speech data that is openly accessible under a Creative Commons CC0 license, establishing it as the most comprehensive public domain corpus suitable for ASR purposes. The Common Voice initiative represents an inclusive and sustainable approach, serving as a viable alternative to both low and high-resource ASR projects (Ardila et al., 2019).

The recording process emphasizes voluntary participation, ensuring that participants are fully informed about the data collection procedure. Users are provided with two options: they can either choose to record anonymously or create accounts. When opting for anonymous recordings, participants relinquish control over their data and forfeit the ability to delete it. Nevertheless, this approach also means that no metadata is linked to their recordings, preserving their privacy. Conversely, users who create accounts benefit from the capability to access and delete their data at any given time. In this scenario, their data is securely stored, ensuring its availability for future use.

This study is firmly committed to upholding ethical principles and responsible practices throughout its execution. An essential aspect of this commitment is the unequivocal assurance that no attempts will be made to identify or extract personal information from the speakers involved in the research. The primary focus remains on the collective and anonymous utilization of speech data for the purposes of conducting the study, reflecting a steadfast dedication to maintaining a respectful and privacy-conscious approach.

To guarantee the highest ethical standards, this study adheres to the guidelines established by the Campus Fryslân Ethical Committee and other applicable regulatory bodies. The Ethics Committee at Campus Fryslân assumes a pivotal role in evaluating and overseeing the ethical aspects of the research, guided by a comprehensive set of codes of ethics. These codes include:

- Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants (in Dutch).
- VSNU Code of Conduct for scientific practice (in Dutch).
- VSNU Code of Conduct for scientific practice (in English).
- ESF European Code of Conduct for research integrity.
- VSNU Code for the use of personal data in scientific research (in Dutch).
- VSNU Code for the use of personal data in scientific research - appendix (in Dutch).
- American Sociological Association Code of Ethics.
- "De AVG en de RUG" (General Data Protection Regulation and the University of Groningen).

Additionally, for research involving human subjects, the evaluation process includes compliance with the following regulations:



- WMO (Dutch).
- Helsinki Declaration.
- CCMO (Central Committee on Research Involving Human Subjects).

Any deviations from the approved study protocol will be immediately reported to the Ethical Committee, ensuring transparency and accountability. The Ethical Committee will thoroughly evaluate any potential ethical concerns arising from the research, paying particular attention to the impact on the elderly Frisian community.

The research holds significant potential for benefiting the Frisian elderly community. The development of a tailored ASR system can address the unique needs of elderly individuals, supporting their communication and access to information. This technology has the capability to enhance the quality of care provided to Frisian-speaking elderly individuals by enabling more accurate and efficient communication between caregivers and patients. This improved understanding of medical needs, preferences, and concerns can lead to personalized and effective care, ultimately improving the overall well-being and satisfaction of the elderly Frisian-speaking population.

To ensure the research aligns with the needs and values of the community, close coordination with local stakeholders, including language policy advisors and community members, will be maintained throughout the study. This collaborative approach will help ensure that the research is conducted in a manner that is sensitive to the community's perspectives and priorities.

Furthermore, the research team is committed to transparently communicating the study's results and implications to the Frisian elderly community. The findings will be disseminated in formats that are accessible, understandable, and useful to the community. By fostering meaningful engagement and empowering the community with knowledge, the research aims to facilitate informed decision-making and active participation.

Transparency, community engagement, and accessible communication are central principles guiding this research. The goal is to develop an ASR system that benefits the Frisian elderly community while minimizing any unforeseen negative consequences, ultimately promoting the well-being and empowerment of the community members.

The code for this research can be accessed on GitHub<sup>1</sup>, and the models are openly available on Hugging Face<sup>2</sup>. The project is designed to be easily replicated, and detailed instructions can be found in the repository. Although the expected results should be similar, it is important to note that there may be some limitations due to the specific environment in which this project was conducted. Specifically, the research was carried out on the high-performance cluster named Hábrók at the University of Groningen.

---

<sup>1</sup><https://github.com/Golesheed/wav2vec2-xls-r-elderly-cv-13-frisian-augmented>

<sup>2</sup><https://huggingface.co/golesheed>



CHAPTER 5 

Results and Discussion

---

This chapter presents the detailed results and discusses their implications. Section 5.1 presents the findings of the experiments described in the Methodology section, while Section 5.2 revisits the research question and hypothesis, providing a reflective analysis of the study and mentioning the limitations.

## 5.1. Results

In this section, the results of two experiments conducted to evaluate the advantages of data augmentation will be presented. Subsection 5.1.1 provides an overview of the results obtained from the baseline experiment. On the other hand, Subsection 5.1.2 focuses on the results derived from the model trained on augmented data.

### 5.1.1 Experiment 1

I conducted my initial experiment by fine-tuning the pre-trained wav2vec 2.0 XLS-R model available on Huggingface<sup>1</sup> using 27 minutes of elderly Frisian speech. The evaluation of this model on the test set highlights the importance of learning cross-lingual representations, particularly for low-resourced languages like Frisian. Since there is no existing model specifically designed for Frisian elderly speech, this model serves as a starting point or baseline. The results show that this model achieves a WER of 47.03% on the development set and 36.21% on the test set, which is an improvement over the previous state-of-the-art model for Frisian ASR as reported by (Yilmaz et al., 2016).

In addition, I further fine-tuned the wav2vec 2.0 XLS-R model using my previous work, which is publicly available on Huggingface<sup>2</sup>. I chose this model as a baseline for the current project because it was already fine-tuned on Frisian and demonstrated better performance, as shown in Table 5.1.

This experiment aimed to determine the best possible outcome by fine-tuning the model on a small dataset, without employing any data augmentations. To achieve this objective, I investigated the influence of different learning rates on the fine-tuning process. As described in Section 4, I followed the range of values specified in Section 4.3 of the research paper by Conneau et al. (2020) to guide my selection of learning rates. I tested with multiple learning rates and the best results can be seen in Table 5.1.

Table 5.1: Results for the wav2vec 2.0 XLS-R models fine-tuned on 27.5 minutes of elderly data. The model fine-tuned on greenwolf/wav2vec2-large-xls-r-1b-frisian serves as the baseline for this project.

Fine-tuned on	WER		Learning rate	Epochs	Training time
	dev	test			
facebook/wav2vec2-xls-r-1b	47.03	36.21	8.00E-05	80	27m
greenwolf/wav2vec2-large-xls-r-1b-frisian	29.77	24.63	8.00E-05	80	27m

<sup>1</sup><https://huggingface.co/facebook/wav2vec2-xls-r-1b>

<sup>2</sup><https://huggingface.co/greenwolf/wav2vec2-large-xls-r-1b-frisian>

### 5.1.2 Experiment 2

Moving forward, I proceeded with generating augmented data using the methods described in Section 4.3. Furthermore, I trained the model on three different datasets. In the first experiment, I used 100% of the clean data along with 20% of data augmented using a band stop filter, 20% with pitch shift, and 20% with Gaussian noise. I conducted multiple trials with various learning rates and determined that a learning rate of  $8e-06$  for 70 epochs produced the best results for the evaluation of the development and test sets.

For the second experiment, I included an additional 20% of data augmented with polarity inversion. In this case, I used the same learning rate of  $8e-06$  and trained the model for 70 epochs. Finally, in the last set of experiments, I incorporated 20% of data augmented using the Gain augmentation method. The model was trained using a learning rate of  $9e-6$  and  $2e-5$ , and the training process lasted for 30 epochs. The results obtained from these experiments are presented in table 5.2.

Table 5.2: Results obtained from fine-tuning the wav2vec 2.0 XLS-R models after augmenting the data. The names used in the table correspond to the explanations provided in reference 4.2.

Data	WER		Hyperparameters		Training time
	dev	test	Learning rate	Epochs	
baseline	29.77	<b>24.63</b>	$8e-05$	80	27m
<b>20p20gn20b20pi20g</b>	27.09	21.26	$9e-06$	30	22m
<b>20p20gn20b20pi20g</b>	26.02	20.82	$2e-05$	30	25m
<b>20p20gn20b</b>	25.75	20.52	$8e-06$	70	30m
<b>20p20gn20b20pi</b>	15.51	<b>19.79</b>	$8e-06$	70	40m

## 5.2. Discussion

I applied my methodology to address my research questions and test my hypothesis. The results of my study show that fine-tuning the XLS-R model with augmented Frisian elderly speech significantly enhances speech recognition performance.

The objective of my research was to improve the wav2vec 2.0 XLS-R model for Frisian elderly speech using basic data augmentation techniques. The wav2vec 2.0 XLS-R model is a multilingual speech recognition model pre-trained with cross-lingual speech representations. Previous experiments have shown that data augmentation methods yield better results than using only clean data.

Based on this knowledge, my hypothesis was that incorporating augmented data with clean data would improve the recognition of elderly Frisian speech using the wav2vec 2.0 XLS-R model. The experimental outcomes support this hypothesis, demonstrating that training exclusively on elderly speech benefits Frisian ASR and improves the recognition performance of baseline models. This validates the contribution of data augmentation techniques to the accuracy and reliability of automatic

recognition systems for elderly Frisian speech. This advancement will lead to the development of more practical and efficient models for Frisian elderly ASR. In Subsection 5.2.1, I will analyze the obtained results and discuss their implications and significance. Additionally, in Subsection 5.2.2, I will examine the potential limitations of the dataset and how they might have affected the outcomes.

### 5.2.1 Reflection

I used the `greenw0lf/wav2vec2-large-xls-r-1b-frisian` model as the baseline since it demonstrated the most favorable results during training. A possible explanation for this success, as noted in Babu et al. (2021), is that the `facebook/wav2vec2-xls-r-1b` model was trained on a mere 15 hours of Frisian data, whereas the `greenw0lf/wav2vec2-large-xls-r-1b-frisian` model underwent additional fine-tuning specifically for Frisian, leading to superior performance.

The usage of augmented data has resulted in improved overall results, primarily due to the increased quantity of data available. The application of various augmentations to the audio has specifically targeted unique features, treating each audio instance as a distinct input. This approach has significantly enhanced the robustness of the ASR system when exposed to new data.

Pitch shifting is a noteworthy augmentation technique that modifies the pitch of the audio by raising or lowering it. This manipulation introduces additional data instances that contribute to the model’s enhanced robustness. Hence, the incorporation of pitch shifting has led to superior outcomes.

Furthermore, the inclusion of Gaussian noise augmentation, which introduces noise into the audio recordings, has proven to be beneficial. Research has demonstrated that adding such noise can enhance the performance of the model. Hence, this augmentation has also played a role in the improved results.

Moreover, the introduction of a band-stop filter aligns the center frequency more closely with the nonlinear characteristics of human hearing in mel space. This alignment ensures that the selected frequency better aligns with the perception of sound by humans.

The polarity inversion technique, which reverses the waveform by multiplying it by  $-1$ , has provided variation and further improved the model’s performance.

However, the gain augmentation, which adjusts the volume of the audio by multiplying it with a random amplitude factor, carries certain limitations. It does not introduce any novel information to the data, and there is a risk of clipping or wrap distortion, potentially leading to the loss of audio segments. As a result, the gain augmentation may not provide significant value and could compromise the integrity of the audio data.

### 5.2.2 Limitations

The analysis of Table 4.1 highlights a notable gender imbalance in the contributed data, with a significant portion coming from female individuals. As a result, the model is primarily trained on such data. However, during testing, the distribution of recordings between males and females in the test set is nearly equal. To address this

discrepancy, applying data augmentations that adjust the pitch, specifically lowering the pitch of female recordings, shows potential for improving the model's performance.

To examine the errors produced by the speech decoder, a thorough inspection of the model's inference on the test dataset is conducted, comparing it to the ground truth labels. The common decoding errors can be summarized into the following categories:

1. Words composed of multiple words are decoded as separate words, such as the model decoding *fierder* as *fier der*.
2. Silent vowels and consonants pronounced by the speaker are not recognized, leading the model to decode *oft* as *at*.
3. Incorrect recognition of diacritics, possibly due to the higher frequency of vowels without diacritics, resulting in the model decoding *ûndersocht* as *undersocht*.
4. Wrong decoding of vowels, where the model predicts *neamt* as *nimt*.
5. Occasional devoicing of consonants by the model, such as decoding *skoalboerd* as *skoalboert*.
6. Incomplete pronunciation of the last word in a sentence, as the end of the sentence is sometimes cut off.

The majority of errors fall into types 1, 3, or 5. Therefore, increasing the amount of data and providing more example sentences can potentially reduce the occurrence of these errors.





CHAPTER 6 

Conclusion

---

## 6.1. Culmination

---

This paper focuses on the impact of data augmentation on elderly Frisian ASR. Building upon a previous framework fine-tuned by myself and a colleague, I explored the enhancements in a target dataset. The architecture employed is based on the wav2vec 2.0 XLS-R, incorporating a multi-layered convolutional feature encoder, a Transformer, and a Quantization module. Specifically, I fine-tuned this model using recordings of elderly individuals aged 65 and above from the Common Voice version 13.0 dataset.

In addition to training the model on the clean dataset, I introduced augmented speech data to the training set. This augmentation aimed to increase the volume of available data and assess the effects on WER.

The results of my experiments validate the hypothesis that data augmentation yields positive effects on the accuracy and reliability of automatic recognition systems for elderly Frisian speech. The relative WER demonstrated an impressive improvement ranging from 15% to 21%.

---

## 6.2. Future Work

---

I intend to gather a new dataset in collaboration with Dr. Lysbeth Jongbloed, who is the Language Policy Advisor for the Province of Fryslân. This dataset will consist of recordings of senior citizens answering daily life questions, which they will send via WhatsApp to me. The data will be collected by asking several prompts (see Appendix B) from the elderly Frisian population. People will be recruited from different locations across the province of Fryslân, as well as from communities outside the country, such as those in Germany or Australia.

This procedure will involve creating a new dataset, which will require obtaining consent from all study participants. Ethical approval has been granted to this project already (see Appendix C for the EA application).

During the informed consent process, participants will receive a detailed explanation of the study's objectives, and procedures, as well as any possible risks and benefits. Participants will have the right to decline or withdraw from the study at any time without consequences. The privacy and confidentiality of all participants will be carefully safeguarded throughout the study. The recordings collected will be anonymized, and securely stored on the researcher's personal device, and access to the data will be limited to the researcher and her supervisory team. The possibility of sharing this dataset with other researchers will be discussed with participants.

I am fully dedicated to complying with the principles and requirements outlined in the General Data Protection Regulation (GDPR). With the utmost respect for individual privacy and data protection, I will ensure that all personal data under my control is collected, processed, and stored legally, transparently, and securely. Furthermore, I will implement appropriate measures to uphold data subject rights, such as the right to access, rectification, erasure, and objection. By adhering to the GDPR's guidelines, I aim to establish a trustworthy and accountable approach to data

handling, fostering a culture of privacy and data protection within my organization.

The data gathered will be further labeled by Frisian-speaking students and will be utilized for fine-tuning the wav2vec 2.0 XLS-R model. To have even more data, I plan to do voice conversion on the Common Voice dataset, to augment all the adult speech to the elderly in order to have more elderly speech data. This is hoped to improve the existing pre-trained model on Frisian and make it work much better with elderly speech data. If data augmentation methods improve the model’s WER, it could be concluded that using voice conversion as a data augmentation method could potentially further lower the WER. Therefore, it is noticeable that using voice conversion to obtain more data from adult speakers to elderly speakers could lead to more speech data. As mentioned in my review of Kim et al. (2021), age-to-age conversion has previously been useful to solve this task. I believe that data augmentation through voice conversion can improve the performance of an elderly Frisian ASR system. The use of similar methods has been proven to improve the accuracy of low-resource language ASR for children’s speech (Shahnawazuddin et al., 2020). This improvement will be achieved by modifying non- or para-linguistic information in the speech. For that purpose, VC is applied to adults’ speech to synthetically generate speech data with acoustic attributes similar to those of elderly speakers. These methods will increase the amount of diverse training data and improve the robustness of the recognition system.

Finally, I will compare the results of that model to the existing one that I acquired from this study. As this will contribute to assessing the effectiveness of the fine-tuned model on a new dataset of elderly speech.

### 6.3. Impact and relevance

---

This study has both short- and long-term effects. The primary objective of this research is to achieve a more precise and robust recognition of elderly speech, which will bring tangible benefits to both the elderly population and the broader Frisian language community.

In the short term, the study aims to improve ASR not only for elderly individuals but also for the Frisian language. Enhanced ASR will lead to more accurate and efficient human-computer interactions, improved accessibility, and an enhanced quality of life for senior citizens who prefer communicating in Frisian.

In the long term, the developed model could play a crucial role in the integration of a healthcare robot used in elderly homes. By incorporating the model as the foundation of a voice assistant robot, I can facilitate natural and seamless communication between senior citizens and computers. This integration will encourage elderly individuals to speak more Frisian while enabling easy communication and faster task execution. A voice assistant robot could serve as a valuable tool to enhance the overall user experience and well-being of senior citizens in their daily lives.

Moreover, the corpus developed for this thesis holds significant potential beyond the immediate research goals. It can be a valuable resource for other researchers, enabling further studies related to underrepresented speaker groups, language preservation, and advancements in the ASR. The availability of this corpus will contribute to

the broader field of speech technology, facilitating the development of more inclusive and effective solutions.

Through these short- and long-term impacts, this study strives to make a meaningful difference in the lives of elderly individuals, promote the usage of the Frisian language, and provide valuable resources for future research endeavors.

# Appendices



APPENDIX A 

## Research Proposal

---

In this Appendix, I will upload my research proposal.

# Evaluation of wav2vec 2.0 Speech Recognition for the Elderly Frisian Population

Golshid Shekoufandeh

April 7, 2023

## Abstract

Automatic Speech Recognition (ASR) is used for transforming human speech into written form, a tool that has become essential in the digitization of daily life – not just for Alexa or Siri, but for the development of mobile tools to assist people in their daily lives. While there are several publicly available ASR models, most are only available in English, and only a few support the Frisian language. None are designed for elderly speakers. This lack of resources means that Frisian elderly speakers are at a double disadvantage when it comes to ASR. As a result, there are significant hurdles to developing technologies for this local community. A promising model for Frisian speech recognition is fine-tuned from the Facebook XLSR Wav2Vec2 model, which achieved a word error rate of 16.25% when trained on the Common Voice dataset. I plan to fine-tune the aforementioned model specifically for Frisian elderly speech in an attempt to achieve a lower word error rate.

**Index Terms:** Frisian, elderly speech, speech recognition, self-supervised learning, wav2vec 2.0, XLSR-53, under-resourced language, low-resourced data, data augmentation



# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Literature review</b>	<b>4</b>
2.1	Introduction and Background . . . . .	4
2.2	Low-Resource Speech Recognition . . . . .	4
2.3	Data Augmentation . . . . .	4
2.4	Conclusion . . . . .	5
<b>3</b>	<b>Research question and hypothesis</b>	<b>6</b>
<b>4</b>	<b>Execution</b>	<b>7</b>
4.1	Methodology and Analysis of the Methods . . . . .	7
4.2	Timeline . . . . .	7
<b>5</b>	<b>Risk mitigation</b>	<b>8</b>
5.1	Risks and contingencies . . . . .	8
<b>6</b>	<b>Ethical issues</b>	<b>9</b>
<b>7</b>	<b>Impact and relevance</b>	<b>9</b>
	<b>References</b>	<b>10</b>

# 1 Introduction

It is not surprising that the world is becoming increasingly digitalized with each passing day, and technology has undoubtedly impacted our modes of communication. The elderly population is no exception to this trend, as technology has affected their communication habits just as much as any other age group. While efforts have been made to facilitate their interaction with technology, these efforts have primarily focused on higher-resourced languages. Unfortunately, there has been very little research conducted on speech recognition of low-resourced languages such as Frisian, and none on recognition of Frisian elderly speech. My thesis will focus on this area and explore ways to address this research gap.

Frysk (West Frisian), a West Germanic language, is predominantly spoken in the province of Fryslân, primarily by individuals of Frisian heritage. Friesland has a population of 643,000. It is estimated that 74% of the population is capable of speaking Frisian, indicating a total of 400,000 Frisian speakers. Currently, slightly over half of the population speaks Frisian at home. Surveys reveal that approximately 94% of the population can comprehend Frisian, 65% can read it, and only 17% can write in the language (Gorter, 2006).

Given the status of Frisian, the demographics of the language speakers, and the increasing importance of speech technology, it is important to ensure that speech recognition systems cater to the needs of Frisian speakers, especially the elderly population, to prevent the further marginalization of the language. Most state-of-the-art speech recognition systems exhibit bias toward individuals whose speech patterns differ from those of norm speakers. These norm speakers are usually adults who are highly educated and speak a standardized language variety as their first language (Zhang et al., 2022). Consequently, most research has focused on developing systems catering to non-elderly, non-child speakers, making elderly or children speech recognition a low-resourced task. This is a concern, especially given that 22% of the population in Friesland is aged above 65, indicating that approximately one-quarter of the population may not have easy access to speech recognition systems that accurately recognize their speech. This, in turn, prevents this population from engaging with smart speakers, many IoT devices, and other such innovations. Over time, this could further delegitimize Frisian and hasten replacement with Dutch while for now it results in a technological gap between those who can use speech technology and those who cannot. To address the gap, I will investigate augmentation techniques to benefit elderly Frisian speech. Data augmentation is a technique that allows us to increase the amount of data by artificially forming new and different data using existing data.

The linguistic situation makes it expensive and challenging to rely on labeled data to implement speech recognition systems since they require large amounts of transcribed speech to achieve high performance. To handle this challenge, using semi-supervised models, like wav2vec 2.0, seem like an appropriate option. Semi-supervised models could be trained with just a small amount of labeled data (as little as one hour), while the rest of the data is unlabeled. Therefore I have chosen to use the wav2vec 2.0 XLSR-53 model, fine-tuned on Frisian. The existing model has achieved a 16.25% Word Error Rate (WER) (de Vries, 2021), representing the baseline for this research.

The rest of this paper is organized as follows. In section 2, I will review the literature I encountered while looking for keywords. Research questions and Hypotheses are presented in Section 3. Methodologies and timelines are shown in section 4. In section 5, I will mention the potential risks of doing this thesis and how to manage them. The ethical considerations are brought up in section 6. The influences of this research are shown in section 7.

## 2 Literature review

To investigate speech recognition systems built for low-resourced languages in more detail, I conducted a literature review. In this section, I will briefly summarize the relevant literature I encountered while researching a combination of the following keywords on *Google Scholar*:

**Speech Recognition** ASR, Automatic Speech Recognition, Speech Recognition, wav2vec 2.0, end-to-end speech recognition, self-supervised learning

**Low-resource** low-resource speech recognition, low-resource automatic speech recognition

**Elderly Speech** Elderly Speech, bias mitigation

**Data Augmentation Methods** Data Augmentation, voice conversion

In this section, I will be synthesizing the literature I encountered in the following subsections. In subsection 2.1, I will be looking at the wav2vec 2.0 model which is the baseline of my model. In subsection 2.2, I will be reviewing how improvement can be achieved for speech recognition in low-resourced languages. Further, in subsection 2.3 different data augmentation methods are mentioned that models could potentially benefit from. Finally, there will be a conclusion to wrap up all the found literature in 2.4.

### 2.1 Introduction and Background

The primary focus of this proposal is to examine state-of-the-art literature that pertains to "wav2vec 2.0". wav2vec 2.0 is a framework presented in Baeovski et al. (2020) for self-supervised learning of speech representations. The framework masks latent representations of the raw waveform and solves a contrastive task over quantized speech representations. One of the corpora used to train and test this framework is Common Voice which was presented by Ardila et al. (2019). In Yi et al. (2020), the authors used a pre-trained model for solving their speech recognition task, and they concluded that wav2vec 2.0 had learned basic acoustic units that can be used to compose various languages.

I will provide a comprehensive list of all the references cited in this context of low-resourced languages and vulnerable individuals, as shown in Table 1.

### 2.2 Low-Resource Speech Recognition

In Coto-Solano et al. (2022), three Automatic Speech Recognition (ASR) systems were trained for Cook Islands Māori (CIM), an indigenous low-resourced language. One is statistical, based on Kaldi toolkit, and the other two were based on Deep Learning, DeepSpeech, and XLSR-wav2vec 2.0. Similarly, Phatthiyaphaibun et al. (2022) fine-tuned a pre-trained XLSR-Wav2Vec2 model to train a Thai Automatic Speech Recognition system using a newer version of CommonVoice corpus.

In May 2022, a Workshop on Language Technology for Equality, Diversity, and Inclusion was held in Ireland. During the workshop, Bharathi et al.; Srinivasan et al.; Suhasini and Bharathi (2022) trained models using pre-trained wav2vec 2.0 to improve speech recognition for Vulnerable Individuals in Tamil, including elderly males, females, and transgender individuals. One key observation they made was that because the pre-trained model used for the system was fine-tuned with the common voice dataset, the model can be trained with one's dataset and used for testing, which can enhance performance.

### 2.3 Data Augmentation

It has been proven that data augmentation could potentially improve the performance of speech recognition systems. Sriram et al. (2022) demonstrated that data augmentation can be utilized with Wav2Vec 2.0. They suggested that implementing their proposed model for cross-lingual representation learning in a similar fashion could considerably benefit languages with restricted speech data. In light of this, it is worth mentioning the augmentation methods that can significantly improve speech recognition for elderly individuals.

For example, Thai et al. (2019) proposed a multistage deep learning approach for low-resource ASR that employs both transfer learning and data augmentation via speech signal distortion and voice conversion.

Similarly, Jin et al. (2022) utilized SVD-based speech spectrum decomposition to derive spectral and temporal subspace representations because the adversarial data augmentation method they introduced necessitates the use of parallel control and recordings of identical spoken content. Furthermore, Shahnawazuddin et al. (2020) utilized voice conversion as a data augmentation method to improve recognition of children’s speech, significantly reducing the word error rate through VC-based out-of-domain data augmentation.

Lastly, Kim et al. (2021) concentrated on improving a recognition system that struggles with recognizing outlier voices, such as the elderly. They proposed age-to-age voice translation using linguistic-coupled information to enhance speech recognition performance for elderly individuals.

## 2.4 Conclusion

As discussed in Section 1, low-resourced languages are at high risk of extinction, primarily due to a lack of transcription. To mitigate this risk, the wav2vec 2.0 model (Baevski et al., 2020) has been proposed as a potential solution.

A review of the literature indicates that training the wav2vec 2.0 model on low-resourced languages can significantly improve their recognition, as demonstrated in studies by (Bharathi et al., 2022; Coto-Solano et al., 2022; Phatthiyaphaibun et al., 2022; Srinivasan et al., 2022; Suhasini & Bharathi, 2022; Yi et al., 2020).

Moreover, it has been observed that data augmentation can also enhance the recognition performance of elderly speech and low-resourced languages, as illustrated in (Jin et al., 2022; Sriram et al., 2022; Kim et al., 2021; Shahnawazuddin et al., 2020; Thai et al., 2019).

Reference	Article
Baevski et al., 2020	Wav2vec 2.0: A framework for self-supervised learning of speech representations
Ardila et al., 2019	Common voice: A massively- multilingual speech corpus
Yi et al., 2020	Applying wav2vec 2.0 to speech recognition in various low-resource languages
Coto-Solano et al., 2022	Development of automatic speech recognition for the documentation of cook islands māori
Phatthiyaphaibun et al., 2022	Thai wav2vec 2.0 with common voice v8
Srinivasan et al., 2022	Speech recognition for vulnerable individuals in Tamil using pre-trained XLSR models
Bharathi et al., 2022	Findings of the shared task on speech recognition for vulnerable individuals in Tamil
Suhasini and Bharathi, 2022	Transformer based approach for speech recognition for vulnerable individuals in Tamil
Sriram et al., 2022	Wav2vec-aug: Improved self-supervised training with limited data
Thai et al., 2019	Synthetic data augmentation for improving low-resource ASR
Jin et al., 2022	Personalized adversarial data augmentation for dysarthric and elderly speech recognition
Shahnawazuddin et al., 2020	Voice conversion-based data augmentation to improve children’s speech recognition in limited data scenario
Kim et al., 2021	Linguistic-coupled age-to-age voice translation to improve speech recognition performance in real environments

Table 1: List of references

### 3 Research question and hypothesis

Based on the literature review, it is noticeable that using voice conversion to obtain more data from adult speakers to elderly speakers could potentially lead to more speech data. As mentioned before, age-to-age conversion has previously been useful to solve this task. To implement this method on elderly Frisian speech, this research question is proposed:

**RQ:** Can using Voice Conversion to augment speech data improve automatic recognition of elderly Frisian speech?

**Hypothesis:** I believe that data augmentation through voice conversion can improve the performance of an elderly Frisian speech recognition system. The use of similar methods has been proven to improve the accuracy of low-resource language speech recognition for children’s speech (Shahnawazuddin et al., 2020). This improvement will be achieved by modifying non- or para-linguistic information in the speech. For that purpose, VC is applied to adults’ speech to synthetically generate speech data with acoustic attributes similar to those of elderly speakers. These methods will increase the amount of diverse training data and improve the robustness of the recognition system.

If the experiment does not support the hypothesis, then the chosen data augmentation technique cannot improve elderly ASR. In that case, other data augmentation techniques should be investigated. There may be other factors that affect the accuracy of ASR systems for elderly speech that are not addressed by these techniques, e.g. age-related changes in speech production, or even microphone placement, etc.

Based on the literature review, it is evident that data augmentation techniques have been successful in improving the accuracy of low-resource language speech recognition systems. However, there is a lack of research on the use of data augmentation for elderly Frisian speech. To address this gap, this research question is presented:

**RQ:** Can data augmentation improve automatic recognition of elderly Frisian speech?

**Hypothesis:** The use of synthetic data augmentation (Thai et al., 2019) and personalized adversarial data augmentation (Jin et al., 2022) will improve the accuracy of low-resource language speech recognition for elderly speech. This improvement will be achieved by creating synthetic speech data based on the unique characteristics of each individual’s speech patterns and by augmenting the existing training data using adversarial data augmentation. These methods will increase the amount of diverse training data and improve the robustness of the recognition system.

If the experiment does not support the hypothesis, then the chosen data augmentation techniques cannot improve elderly ASR. In that case, other data augmentation techniques should and will be investigated. There may be other factors that affect the accuracy of ASR systems for elderly speech that are not addressed by these techniques, e.g. age-related changes in speech production, or even microphone placement, etc.

## 4 Execution

After mentioning what has been done previously by other researchers and defining the research question and hypothesis, In this section, I will be elaborating on the methodologies of how I will be doing this research. I will explain the model which I will be fine-tuning for elderly speech. It is important to mention that this model was not fine-tuned for this population prior to this research.

For the sake of this proposal, I will mainly try to target and answer my first research question.

### 4.1 Methodology and Analysis of the Methods

I will begin with testing and refining the pre-trained model. The model is wietse/v/wav2vec2-large-xlsr-53-frisian. This pre-trained model is based on facebook/wav2vec2-large-xlsr-53 and was fine-tuned on Frisian using the Common Voice dataset. I will use the Common Voice Corpus 12.0, which includes 50 hours of speech. This corpus is obtained through crowdsourcing and audio validation (Ardila et al., 2019). Contributors can use either the Common Voice website to record their voice while reading sentences displayed on the screen. The recordings undergo a verification process by other contributors through a straightforward voting system. Additionally, I will take subsets (the only decades existing in the test dataset are speakers in the 60s and 80s) from the dataset into various decades ranging from 60+ (Wilpon & Jacobsen, 1996) to evaluate the pre-trained Frisian model's performance. This will indicate how well the pre-trained model performs with elderly speech.

Once I assess the model's proficiency with read speech, I intend to gather a new dataset in collaboration with Dr. Lysbeth Jongbloed who is a Language Policy Advisor for the Province of Fryslân. This dataset will consist of recordings of senior citizens answering daily life questions, which they will send via WhatsApp to me. This unlabeled data will be utilized for fine-tuning the wav2vec 2.0-xlsr-53 model. For the labeled data, I plan to do voice conversion on the Common Voice dataset, to augment all the adult speech to the elderly in order to have more elderly speech data. This is hoped to improve the existing pre-trained model on Frisian and make it work much better with elderly speech data. Finally, I will compare the results of this model to the existing one. As this will contribute to assessing the effectiveness of the fine-tuned model on a new dataset of elderly speech.

### 4.2 Timeline

In this section, I will be showing a Gantt Chart in figure 1 of how the thesis process.

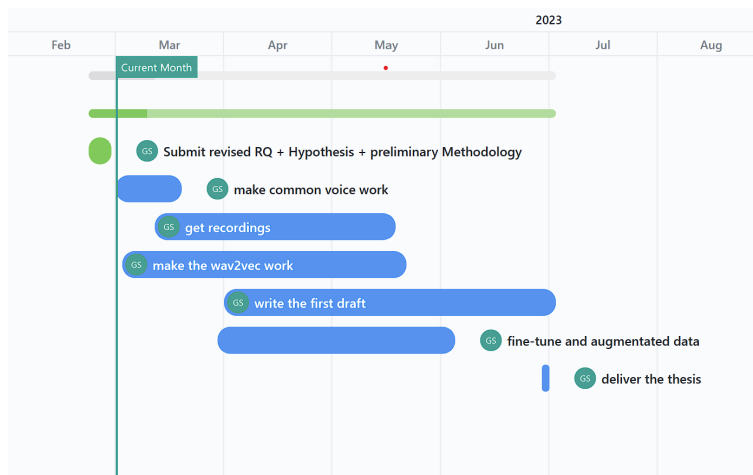


Figure 1: Gantt chart of this thesis

## 5 Risk mitigation

In this section, a brief description of the risks and ways to mitigate them will be presented.

### 5.1 Risks and contingencies

Introducing a recognition system for Frisian elderly speech comes with unique obstacles and hazards. These include limited data availability, dialectical variations, age-related speech changes, and ethical considerations.

To elaborate further, Frisian is a relatively minor language spoken by a small population, implying that data availability may be scarce for training an accurate and reliable speech recognition system specifically for Frisian. However, with the right data augmentation methods, this issue can be addressed.

Additionally, Frisian has various dialects, and elderly speakers may use a different dialect from the one utilized to train the recognition system, leading to decreased accuracy, reliability, and user frustration. To address this, gathering more data and training the model on this data is crucial but might not be doable in a short time as it needs the mentioned data to be gathered separately.

Furthermore, elderly individuals may experience speech changes due to age-related factors such as hearing loss, reduced lung capacity, or dental problems, making it challenging for the recognition system to accurately capture and interpret their speech, possibly resulting in more errors and user dissatisfaction. Like with any recognition system, there are ethical concerns related to privacy, security, and data protection that must be addressed to safeguard elderly users.

Additionally, it is fitting to acknowledge that potential challenges may arise during the implementation of the model and that any necessary further research and support from supervisors can be sought to address them.

Despite the challenges, implementing a recognition system for Frisian elderly speech could significantly improve communication, socialization, and quality of life for this population. Careful planning, design, and testing can help to mitigate potential risks and ensure that the system is effective, reliable, and well-received by its users. Table 2 shows the possible risk and contingency.

Risks	Very likely	Likely	Not likely
Very severe		Cannot train wav2vec 2.0 model	ethical considerations
Severe		various dialects	
Not severe		limited data availability	

Table 2: Risk and Contingency table

## 6 Ethical issues

The study will commence by utilizing the Common Voice Corpus 12.0 dataset, which was gathered and validated through Mozilla’s Common Voice initiative. This dataset contains speech data that is openly available under a Creative Commons CC0 license, making it the most comprehensive public domain corpus available for Automatic Speech Recognition. Common Voice is an open, sustainable alternative to other low/high resource projects(Ardila et al., 2019).

The second part of the study will involve creating a new dataset, which will require obtaining consent from all study participants. During the informed consent process, participants will receive a detailed explanation of the study’s objectives, and procedures, as well as any possible risks and benefits. Participants will have the right to decline or withdraw from the study at any time without consequences. The privacy and confidentiality of all participants will be carefully safeguarded throughout the study. The recordings collected will be anonymized, and securely stored on the researcher’s personal device, and access to the data will be limited to the researcher and her supervisory team. The possibility of sharing this dataset with other researchers will be discussed with participants. Any personal information obtained from participants will be handled in compliance with GDPR. The ethical guidelines set by the CF Ethical Committee and other applicable regulatory bodies will be strictly followed. Any deviation from the approved study protocol will be promptly reported to the Ethical Committee. The Committee will also consider any potential ethical concerns, including the impact of the research on the elderly Frisian community. The study will be closely coordinated with local stakeholders, such as language policy advisors and community members, to ensure that the research is in line with the needs and values of the community. Additionally, the research team will provide accessible and transparent communication of the study’s results and implications to the community. The study aims to develop an ASR system that benefits the Frisian elderly community while minimizing unforeseen negative consequences.

## 7 Impact and relevance

This study has both short- and long-term effects. The primary objective of this research is to achieve a more precise and robust recognition of elderly speech. In the short run, the study could improve speech recognition not only for the people of the age speaker group but also for Frisian as a whole. In the long run, the model could be integrated into a healthcare robot utilized in elderly homes to enhance the communication between senior citizens and computers. In doing so, people of age will be encouraged to speak more Frisian, a language that they are more comfortable with, while at the same time interacting with computers in such a way that communication is easy and tasks are performed much faster. Moreover, the corpus developed for this thesis could be beneficial to other researchers in the future and can be used further for studies related to speaker groups with less digital representation.



## References

- Bharathi, B., Chakravarthi, B. R., Cn, S., Sripriya, N., Pandian, A., & Valli, S. (2022). Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 339–345.
- Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., Koka'ua, L., Tanveer, S., & Feldman, I. (2022). Development of automatic speech recognition for the documentation of Cook Islands Māori.
- Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., & Liu, X. (2022). Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition. *arXiv preprint arXiv:2205.06445*.
- Phatthiyaphaibun, W., Chaksangchaichot, C., Limkonchotiwat, P., Chuangsuwanich, E., & Nutanong, S. (2022). Thai Wav2Vec2. 0 with CommonVoice V8. *arXiv preprint arXiv:2208.04799*.
- Srinivasan, D., Bharathi, B., Durairaj, T., et al. (2022). SSNCSE\_NLP@ LT-EDI-ACL2022: Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR models. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 317–320.
- Sriram, A., Auli, M., & Baevski, A. (2022). Wav2Vec-Aug: Improved self-supervised training with limited data. *arXiv preprint arXiv:2206.13654*.
- Suhasini, S., & Bharathi, B. (2022). SUH-ASR@ LT-EDI-ACL2022: Transformer based Approach for Speech Recognition for Vulnerable Individuals in Tamil. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 177–182.
- Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2022*, 3168–3172.
- de Vries, W. (2021). WIETSEDV/WAV2VEC2-large-XLSR-53-frisian · hugging face. <https://huggingface.co/wietsedv/wav2vec2-large-xlsr-53-frisian>
- Kim, J.-W., Yoon, H., & Jung, H.-Y. (2021). Linguistic-coupled age-to-age voice translation to improve speech recognition performance in real environments. *IEEE Access*, 9, 136476–136486.
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.
- Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., & Ahmad, W. (2020). Voice Conversion Based Data Augmentation to Improve Children’s Speech Recognition in Limited Data Scenario. *Interspeech*, 4382–4386.
- Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv:2012.12121*.
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- Thai, B., Jimerson, R., Arcoraci, D., Prud’hommeaux, E., & Ptucha, R. (2019). Synthetic data augmentation for improving low-resource ASR. *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, 1–9.
- Gorter, D. (2006). Gorter, D.(2006) La Llengua Frison en los Paisos Bajos [in Catalan: The Frisian language in the Netherlands], special issue on ‘Europa parla. Llengües no romàniques minoritzades d’Europa’, in Anuari. Revista de recerca humanística i científica, 2006: XVII, Universitat Jaume I, ISSN: 1130-4235, pp 51-63.[this is the original English version].
- Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, 1, 349–352.



## APPENDIX B

### Prompts

---

Here are the prompts, asked by us from the elderly. The elderly were asked to answer as many questions and elaborately as possible.

- *Wat wie it hichtepunt fan jo dei oant no ta? Hat er ien by jo op besite west? Familje, freonen? Binne jo sels noch fuort west?*
- *Wat ha jo hjoed iten? Hawwe jo sels it iten klearmakke? Wat hawwe jo dien?*
- *Hoe is it waar hjoed en wat is de waarsferwachting fan moarn? Hoe hat it waar jo plannen beynfloede/beynfloed it waar jo plannen fan dizze wike?*
- *Wat is jo favorite seizoen? Wêrom?*
- *Wat ha jo hjoed iten?*
- *Wat wie jo earste baan? Kinne jo jo noch dingen fan jo wurkplak werinnerje? Jo baas? Jo kollega's?*
- *Fertel my ris wêr't jo opgroeid binne. Hoe seach jo buert derút? Hoe wie jo bernetiid? Tinke jo dat bern it tsjintwurdich beter ha as yn jo tiid? Wêrom?*
- *Wat foar dei is it hjoed? Hoelet is it no? Hoe let waard jo wekker? Wannear ha jo jo moarnsiten, middeisiten en jûnsiten hân?*
- *Hoe sille jo nei Harns ta as jo dêrhinne wolle? Hokker trein nimme jo dan?*
- *Wat wie jo earste húsdier? Kinne jo wat oantinkens diele oer it libben mei dat húsdier?*
- *Wat wie jo meast resinte húsdier?*
- *Fan hokker hobbys genietsje jo? Wat makket dy hobby moai foar jo?*
- *Hoe is it Frysk tsjintwurdich oars as yn jo bernetiid?*

**Prompts in English:**

- What was the highlight of your day so far? Did anyone visit you? Family, friends?
- What did you eat today? How was the food prepared? What's the weather like today and what is the forecast for tomorrow? How has the weather affected/- does the weather affect your plans this week?
- What's your favorite season? Why?
- What did you eat today?
- What was your first job? Do you remember any of the details about your workspace? Your boss? Your colleagues?
- Tell me about where you grew up. What was your neighborhood like? How was your childhood? Do you think children today are better off now than in your time? Why?
- What day is today? What time is it right now? What time did you wake up? When did you eat your breakfast, lunch, and dinner?
- How will you go to Harlingen if you want to go now? Which train will you take?
- What was your first pet? Can you share some memories about life with that pet?
- What was your most recent pet?
- What hobbies do you enjoy? What makes them enjoyable for you?
- How is Frisian different today than in your childhood?



## Ethical Approval Application

This study targeted vulnerable individuals. Therefore, I had to ask for ethical approval from the Campus Fryslân Ethics Committee. This is the form sent to them.





**Section B**

**Research Proposal Template**

<b>Project Title</b>	Evaluation of Wav2vec2.0 Speech Recognition for the Elderly Frisian population
List of any sources of funding or other research partners involved	Provinsje Fryslân
Is this proposal associated with another research study?	no
Expected dates of commencement and completion (fieldwork)	The project will go on for a month as soon as I get approval.
Abstract of the proposal	Automatic Speech Recognition (ASR) is used for transforming human speech into written form, a tool that has become essential in the digitization of daily life -- not just for Alexa or Siri, but for the development of mobile tools to assist people in their daily lives. While publicly available ASR models have been released in HuggingFace, most of these are only available in English, and very few are available in Frisian, with none specifically designed for elderly individuals. This lack of resources means that Frisian elderly speakers are at a double disadvantage when it comes to ASR. As a result, there are significant hurdles to developing technologies for this local community. One of the best models for Frisian speech recognition is fine-tuned from the Facebook XLSR Wav2Vec2 model It was trained using the Common Voice Dataset, and achieves a word error rate of 16.25%. In this paper, we aim to fine-tune this model specifically for elderly speech to obtain a lower word error rate.
Rationale and background of the proposed study	Most of the state-of-the-art speech recognition systems exhibit bias toward individuals whose speech patterns differ from those of norm speakers. These norm speakers are usually adults who are highly educated and speak a standardized language variety as their first language ( <a href="#">Zhang et al. 2022</a> ). Consequently, most of the research has focused on developing systems that cater to non-elderly, non-child speakers, making it a low-resource area. This is a concern, especially given that 22% of the <a href="#">population</a> in Friesland is aged above 65, indicating that approximately one-quarter of the population may not have easy access to speech recognition systems that accurately recognize their speech. This, in turn, prevents this population from engaging with smart speakers, many IoT devices, and other such innovations. Over time, this could further delegitimize Frisian and hasten replacement with Dutch while for now it results in a technological gap between those who can use speech technology and those who cannot. To address this, I will investigate techniques to augment existing data sets to the benefit of elderly Frisian speech.
Research question, aims and objectives	Can data augmentation improve automatic recognition of elderly Frisian speech?
Hypothesis	The use of synthetic data augmentation ( <a href="#">Thai et al. 2019</a> ) and personalized adversarial data augmentation ( <a href="#">Jin et al. 2022</a> ) will improve the accuracy of low-resource language speech recognition for elderly speech. This improvement will be achieved by creating synthetic speech data based on the unique characteristics of each individual's speech patterns and augmenting the existing training data using adversarial data augmentation. These methods will increase the amount of diverse training data and improve the robustness of the recognition system.

	<p>If the experiment does not support the hypothesis, then the chosen data augmentation techniques cannot improve elderly ASR. In that case, other data augmentation techniques should and will be investigated.</p>
<p>Outline of the research design and analysis</p>	<p>I will begin with testing and refining the pre-trained model, The model is <a href="#">wietsedv/wav2vec2-large-xlsr-53-frisian</a>. This pre-trained model is fine-tuned of <a href="#">facebook/wav2vec2-large-xlsr-53</a> on Frisian using the Common Voice dataset. I will use the Common Voice Corpus 12.0, which includes 150 hours of speech and involves 1,422 speakers. This corpus is obtained through crowdsourcing and audio validation (<a href="#">Ardila et al. 2019</a>). Contributors can use either the Common Voice website or iPhone app to record their voice while reading sentences displayed on the screen. The recordings undergo a verification process by other contributors through a straightforward voting system. Additionally, I will take subsets (the only decades existing in the test dataset are in the 60s and 80s) from the dataset into various decades ranging from 60+ (<a href="#">Wilpon et al. 2016</a>) to evaluate the pre-trained Frisian model's performance.</p> <p>Once I assess the model's proficiency with read speech, I intend to gather a new dataset in collaboration with Dr. <a href="#">Lysbeth Jongbloed</a> who is a Language Policy Advisor for the Province of Fryslân. This dataset will consist of recordings of senior citizens answering daily life questions, which they will send via WhatsApp to the author. This unlabeled data will be utilized for pre-training and fine-tuning the <a href="#">wav2vec2.0-xlsr-53</a> model. Further, I will implement data augmentation methods that have been previously mentioned to improve the fine-tuned model on elderly Frisian speech. Finally, I will compare the results of this model to the existing one.</p>
<p>When research involves access to human participants outline fully where and how they will be recruited, inclusion and exclusion criteria and the exact role of any gatekeepers involved</p>	<p>Recruitment: Participants will be recruited through collaboration with local community centers and senior care facilities in the Frisian region. Digital announcements advertising the study will be shared on relevant social networks. Interested individuals will be asked to contact the research team via email. The research team will also leverage existing connections with local organizations and community leaders to reach out to potential participants.</p> <p>Inclusion criteria:</p> <ul style="list-style-type: none"> <li>- Participants must be 60 years of age or older.</li> <li>- Participants self-identify as Frisian speakers.</li> <li>- Participants must have access to a smartphone with WhatsApp installed.</li> </ul> <p>Exclusion criteria:</p> <ul style="list-style-type: none"> <li>- Participants with a history of speech or hearing disorders that could affect speech recognition</li> <li>- Participants with cognitive impairment that may impact their ability to complete the study tasks will be excluded from the study.</li> </ul> <p>Gatekeepers: Language Policy Advisor Dr. <a href="#">Lysbeth Jongbloed</a> will serve as a gatekeeper for this study. Dr. Jongbloed will assist in recruitment efforts by providing access to local community centers and senior care facilities, and will also serve as a liaison between the research team and potential participants. She will help to ensure that all participants meet the inclusion criteria and will assist with scheduling study sessions.</p>

Any additional information	
----------------------------	--

<b>Section C</b>
------------------

<b>Please answer the following questions (Y/N)</b>	<b>(Y/N)</b>
1. Will any non-anonymized and/or personalized data be generated and/or stored?	N
2. Will your project involve any of the following?	Photographing Participants N
	Audio Recordings Y
	Video Recordings N
3. Does this research pose any risk of physical danger to the researcher?	N
4. Does this research pose any risk of mental harm to the researcher?	N
5. Will you give the potential participants a reasonable period of time to consider participation?	Y
6. Does your study involve any of the following?	People who are, have been, or are likely to become your clients, students, or clients of the School N
	Patients N
	Children (under 18 years of age) N
	People with intellectual or communication difficulties N
	People in custody N
	People involved in illegal activities N
	People belonging to a vulnerable group, other than those listed above Y
	People for whom English / Dutch is not their first language Y
7. Is there any realistic risk of any participants experiencing a detriment to their interests as a result of participation?	N

8. Will you have access to documents containing sensitive data about living individuals? If yes, will you gain the consent of the individuals concerned?	N
9. Has this research application or any application of a similar nature connected to this research project been refused ethical approval by another review committee of the College or any external organization?	N

If you answered yes to any of the above questions please explain with reference to the number of each question, how the identified potential research ethics issue will be handled. If there are any other potential ethical issues that you think the Committee should consider please explain them here. *There is an obligation on the lead researcher/supervisor to consider here any issues with ethical implications not clearly covered above.*

- Informed consent will be obtained from all participants involved in the study. The informed consent process will include an explanation of the study's objectives, procedures, and any potential risks and benefits. Participants will have the option to decline participation or withdraw from the study at any time without penalty.
- Privacy and confidentiality of the participants will be strictly maintained throughout the study. All recordings collected will be de-identified and stored securely on the researcher's personal device, and access to the data will be limited to the researcher and her supervisory team. The possibility of making this dataset available to other researchers will be considered after discussion with participants. Any personal information obtained from the participants will be handled in accordance with GDPR.
- I will adhere to the ethical guidelines set forth by the CF Ethical Committee and any other applicable regulatory bodies. Any deviations from the approved study protocol will be reported to the Ethical Committee promptly.
- Other potential ethical issues that the Committee may wish to consider include the potential impact of the research on the elderly Frisian community. While the study aims to develop an ASR system that will benefit this population, there is a possibility that the technology may have unforeseen consequences. To mitigate this risk, the study will be conducted in close collaboration with local stakeholders, including language policy advisors and community members, to ensure that the research is aligned with the needs and values of the community. Additionally, the research team will communicate the study's results and implications to the community in an accessible and transparent manner.


### Appendices

e.g. participant information sheet, consent form, interview questions, survey

- A consent form has been made and it will be sent to the volunteers.
- Participant information sheet is limited to questions about year/decade of birth,
  - In what year were you born? Or, if you prefer not to specify, please indicate the decade in which you were born (e.g. 1940s, 1950s, 1960s ...)
  - What is your gender?
  - How often do you speak Frisian?
- The prompts that the participants will answer (in Frisian). The main point of the prompts is to elicit spontaneous speech:
  - What was the highlight of your day so far? Did anyone visit you? Family, friends?
  - What did you eat today? How was the food prepared?
  - What was your first pet? Can you share some memories about life with that pet?
  - What was your most recent pet?
  - What hobbies do you enjoy? What makes them enjoyable for you?
  - How is Frisian different today than in your childhood?
  - What's the weather like today and what is the forecast for tomorrow?
  - How has the weather affected/does the weather affect your plans this week?
  - What's your favorite season? Why?
  - 
  - What was your first job? Do you remember any of the details about your workspace? Your boss? Your colleagues?
  - Tell me about where you grew up. What was your neighborhood like? How was your childhood? Do you think children today are better off now than in your time? Why?
  - What day is today? What time is it right now? What time did you wake up? When did you eat your breakfast, lunch, and dinner?
  - How will you go to Harlingen if you want to go now? Which train would you take?


**Section D**

I confirm that this application provides a complete and accurate account of the research I propose to conduct in this context, including my assessment of the ethical ramifications. I undertake to return for additional ethical approval should any design changes warrant it.

Signed:  \_\_\_\_\_ Date: March 15, 2023  
Lead Researcher / Student

**Supervisor's Declaration** (where applicable)

As the supervisor for this project, I confirm that I believe that all research ethical issues have been dealt with in accordance with School policy and the research ethics guidelines of the relevant professional organization. I undertake to continue to review this project and ensure that ethical principles are upheld at every stage.

Signed:  \_\_\_\_\_ Date: March 15, 2023  
Supervisor

*There is an obligation on the lead researcher/supervisor to bring to the attention of the REAC any issues with ethical implications not clearly covered above.*



## Bibliography

---

- Anidjara, O. H., Marbele, R., Abdallaa, N., Bigona, N., Myaraa, B., & Yozeviticha, R. (2023). Augmented Wav2Vec 2.0: ASR Improvement Using Data Augmentation for Under-Represented Languages. <https://github.com/neryabigon/wav2vec2-large-xlsr-53-finetuning/tree/main>.
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.
- Balan, & Shekoufandeh (2023). greenwolf/wav2vec2-large-xls-r-1b-frisian. URL <https://huggingface.co/greenwolf/wav2vec2-large-xls-r-1b-frisian>
- Bharathi, B., Chakravarthi, B. R., Cn, S., Sripriya, N., Pandian, A., & Valli, S. (2022). Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, (pp. 339–345).
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 8789–8797).
- Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 8188–8197).
- Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

- Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., Koka'ua, L., Tanveer, S., & Feldman, I. (2022). Development of automatic speech recognition for the documentation of Cook Islands Māori.
- Eide, E., & Gish, H. (1996). A parametric approach to vocal tract length normalization. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1, (pp. 346–348). IEEE.
- El Helou, M., & Süsstrunk, S. (2020). Blind universal Bayesian image denoising with Gaussian noise level learning. *IEEE Transactions on Image Processing*, *29*, 4885–4897.
- Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., & Velimirović, M. (2020). SPICE: Self-Supervised Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1118–1128.
- Gorter, D. (2006). Gorter, D.(2006) La Llingua Frison en los Paisos Bajos [in Catalan: The Frisian language in the Netherlands], special issue on ‘Europa parla. Llengües no romàniques minoritzades d’Europa’, in Anuari. Revista de recerca humanística i científica, 2006: XVII, Universitat Jaume I, ISSN: 1130-4235, pp 51-63.[this is the original English version].
- Graves, A. (2012). Connectionist temporal classification. *Supervised sequence labelling with recurrent neural networks*, (pp. 61–93).
- Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., & Liu, X. (2022). Personalized Adversarial Data Augmentation for Dysarthric and Elderly Speech Recognition. *arXiv preprint arXiv:2205.06445*.
- Jordal, I. (2023). Audiomentations. <https://github.com/iver56/audiomentations>.
- Kaur, P., Wang, Q., & Shi, W. (2022). Fall detection from audios with audio transformers. *Smart Health*, *26*, 100340.
- Kim, J.-W., Yoon, H., & Jung, H.-Y. (2021). Linguistic-coupled age-to-age voice translation to improve speech recognition performance in real environments. *IEEE Access*, *9*, 136476–136486.
- Li, Y. A., Zare, A., & Mesgarani, N. (2021). StarGANv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. *arXiv preprint arXiv:2107.10394*.
- Phatthiyaphaibun, W., Chaksangchaichot, C., Limkonchotiwat, P., Chuangsuwanich, E., & Nutanong, S. (2022). Thai Wav2Vec2. 0 with CommonVoice V8. *arXiv preprint arXiv:2208.04799*.
- Robinson-Jones, C., & Scarse, Y. (2022). Report on the west frisian language (language technology support of europe’s languages in 2020/2021-european language equality project).



- 
- Roonizi, A. K., & Jutten, C. (2021). Band-stop smoothing filter design. *IEEE Transactions on Signal Processing*, 69, 1797–1810.
- San, N., Bartelds, M., Billings, B., de Falco, E., Feriza, H., Safri, J., Sahrozi, W., Foley, B., McDonnell, B., & Jurafsky, D. (2023). Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions. *arXiv preprint arXiv:2302.04975*.
- Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., & Ahmad, W. (2020). Voice Conversion Based Data Augmentation to Improve Children’s Speech Recognition in Limited Data Scenario. In *Interspeech*, (pp. 4382–4386).
- Srinivasan, D., Bharathi, B., Durairaj, T., et al. (2022). SSNCSE\_NLP@ LT-EDI-ACL2022: Speech Recognition for Vulnerable Individuals in Tamil using pre-trained XLSR models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, (pp. 317–320).
- Sriram, A., Auli, M., & Baevski, A. (2022). Wav2Vec-Aug: Improved self-supervised training with limited data. *arXiv preprint arXiv:2206.13654*.
- Suhasini, S., & Bharathi, B. (2022). SUH\_ASR@ LT-EDI-ACL2022: Transformer based Approach for Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, (pp. 177–182).
- Thai, B., Jimerson, R., Arcoraci, D., Prud’hommeaux, E., & Ptucha, R. (2019). Synthetic data augmentation for improving low-resource ASR. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, (pp. 1–9). IEEE.
- Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2020). Applying wav2vec2.0 to speech recognition in various low-resource languages. *arXiv:2012.12121*.
- Yilmaz, E., Andringa, M., Kingma, S., Dijkstra, J., Kuip, F., Velde, H., Kampstra, F., Algra, J., Heuvel, H., & van Leeuwen, D. A. (2016). A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research.
- Yilmaz, E., van den Heuvel, H., & Van Leeuwen, D. (2016). Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science*, 81, 159–166.
- Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, (pp. 3168–3172).
- Zheng, G., Xiao, Y., Gong, K., Zhou, P., Liang, X., & Lin, L. (2021). Wav-bert: Cooperative acoustic and linguistic representation learning for low-resource speech recognition. *arXiv preprint arXiv:2109.09161*.







# Evaluation of wav2vec 2.0 Speech Recognition for the Elderly Frisian Population

Automatic Speech Recognition (ASR) converts speech into text. It has become crucial in daily life, as evident through the utility of virtual assistants like Alexa and Siri and other tools that help people. Most publicly available ASR models are designed for the English language. Only a few support Frisian and under-resourced Germanic language. Moreover, none of these models are tailored explicitly for elderly speakers. The lack of adequate ASR resources for the Frisian language poses an intersectional disadvantage for elderly speakers, resulting in significant challenges in developing technologies to address the needs of this community. To address this gap, increasing the availability of training data is necessary. In this study, I propose using data augmentation techniques to augment elderly audio recordings. These augmented datasets will be used to train the wav2vec 2.0 XLS-R model, which has shown promise in Frisian ASR. My co-developed model, fine-tuned from the Facebook XLS-R Wav2Vec2 model, achieved a word error rate (WER) of 15.35% when trained on the Common Voice dataset. The main objective of this research is to investigate the effect of fine-tuning the model using augmented elderly speech data tailored explicitly for Frisian elderly speakers. By integrating this dataset, I expanded the collection of recorded speeches from elderly Frisian individuals, leading to a remarkable 20% improvement in relative WER for Frisian elderly ASR. This study makes a valuable contribution towards tackling the technological hurdles encountered by the local Frisian community. Furthermore, it emphasizes the significance of advancing ASR technologies for languages with limited resources and specific demographic groups. Apart from addressing the research objectives, this study offers essential contextual information, underscores the study's importance, and recognizes the broader implications for ASR research in low-resource languages and elderly ASR.

**Index Terms:** Frisian, elderly speech, speech recognition, self-supervised learning, wav2vec 2.0, XLSR-53, under-resourced language, low-resourced data, data augmentation