



**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

**Transfer Learning for Low-Resource Languages:  
Analysing the Effects of Bias and Pre-training  
Language**

**Caoilfhionn Ní Dheoráin, BA (Mod)**

**A Dissertation**

Presented to the University of Dublin, Trinity College  
in partial fulfilment of the requirements for the degree of

**Master in Computer Science (MCS)**

Supervisor: Prof. Anthony Ventresque

April 2024

# Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

---

Caoilfhionn Ní Dheoráin

April 17, 2024

## Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

---

Caoilfhionn Ní Dheoráin

April 17, 2024

# Transfer Learning for Low-Resource Languages: Analysing the Effects of Bias and Pre-training Language

Caoilfhionn Ní Dheoráin, Master in Computer Science (MCS)  
University of Dublin, Trinity College, 2024

Supervisor: Prof. Anthony Ventresque

This dissertation investigates the challenges faced by the Irish language, biases within Irish open source datasets, the efficacy of gloss analysis in determining datasets to improve Automatic Speech Recognition (ASR) systems performance, and the impact of transfer learning on ASR performance. Research on this topic reveals bias against low-resource languages specifically related to dialects and a lack of technological support due to limited research and funding, as well as underlying bias that exists across ASR systems for any language and the damaging effects of them. Analysis of the Mozilla Common Voice datasets Mozilla Corporation (2021) was performed to uncover bias and to link the performance of models trained on these datasets to the data. Under-representation of certain groups was found, particularly women and those over the age of 60. Models were fine-tuned on a set of datasets before being fine-tuned again on Irish to prove that transfer learning on selected languages is a promising approach to enhance ASR performance, with findings indicating a 9.5% improvement in Word Error Rate (WER) through pre-training on English data before fine-tuning on Irish. I also found that bias in the models could not be solely linked back to the metadata from the datasets, meaning that a more in-depth investigation must be done into where the discrimination in the model performance is coming from. Surprisingly, the study finds a lower direct correlation between dataset size and model performance than expected, highlighting the importance of dataset selection. Gloss analysis offers some insights into suitable datasets for pre-training but can't distinguish a dataset that will boost the performance of a model from one that will negatively impact it.

# Acknowledgments

I wish to acknowledge with thanks, the people who gave me enormous help during my dissertation. Principally, my supervisor Prof. Anthony Ventresque, who has given me the chance to continue on with research, and the person who worked most with me and gave me encouragement and direction, Dr. Ellen Rushe. I would also like to extend my gratitude to my brother Seán Ó Deoráin for your listening ear and your wise words, and to my family who I love, for your patience and unconditional support.

CAOILFHIONN NÍ DHEORÁIN

*University of Dublin, Trinity College*

*April 2024*

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgments</b>	<b>iv</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Objectives . . . . .	3
<b>Chapter 2 State of the Art</b>	<b>5</b>
2.1 Background . . . . .	5
2.1.1 Low-Resource Languages . . . . .	5
2.1.2 Transfer Learning . . . . .	6
2.1.3 Wav2Vec . . . . .	6
2.2 Related Work . . . . .	9
2.2.1 Bias in Automatic Speech Recognition (ASR) . . . . .	9
2.2.2 Specific Challenges For Low-Resource Languages . . . . .	11
2.2.3 Challenges of Dialects . . . . .	13
2.2.4 Challenges of Minority Speech Patterns . . . . .	14
2.2.5 Assumptions About Our Data . . . . .	14
<b>Chapter 3 Methodology</b>	<b>17</b>
3.1 Data . . . . .	17
3.2 Fine-Tuned Models . . . . .	18

3.3	Fine-Tuning Models on Irish . . . . .	19
<b>Chapter 4 Experimental Setup</b>		<b>21</b>
4.1	Dataset . . . . .	22
4.1.1	Common Voice . . . . .	22
4.2	Data Analysis of the Irish Common Voice Dataset . . . . .	23
4.3	Results of Investigation into Data Analysis Results of the Irish Common Voice Dataset . . . . .	32
4.4	Text Analysis of the Irish Common Voice Dataset . . . . .	34
4.4.1	Part-Of-Speech(POS) tagging . . . . .	35
4.4.2	Word Distribution . . . . .	37
4.5	Results of Investigation into Text Analysis Results of the Irish Common Voice Dataset . . . . .	39
<b>Chapter 5 Evaluation</b>		<b>41</b>
5.1	Metrics . . . . .	42
5.2	Models Performance Results . . . . .	43
5.3	Similarity of Datasets . . . . .	46
5.3.1	Cosine Similarity . . . . .	46
5.3.2	Distribution of Part-Of-Speech (POS) Tags Comparison . . . . .	47
5.3.3	Distribution of Word Frequency Comparison . . . . .	49
5.3.4	Results of the Similarity of Datasets . . . . .	50
5.4	Bias Analysis of Results . . . . .	51
<b>Chapter 6 Conclusions &amp; Future Work</b>		<b>60</b>
6.1	Future Work . . . . .	62
<b>Bibliography</b>		<b>71</b>

# List of Tables

4.1	Metadata Tags for Irish Dataset . . . . .	22
4.2	Age groups of the Irish Dataset vs Irish speaking population . . . . .	32
4.3	Female age groups of the Irish dataset Vs Irish speaking population . . . . .	33
4.4	Male age groups of the Irish dataset Vs Irish speaking population . . . . .	33
4.5	Most Common Part-of-speech tags , meanings, and examples (In order of frequency) . . . . .	36
4.6	Least Common Part-of-speech tags, meanings, and examples . . . . .	36
5.1	Comparison of Model Performance . . . . .	44
5.2	Correlation Matrix of Performance Metrics . . . . .	45
5.3	Cosine Similarity Scores . . . . .	50
5.4	Best Performing Models for Each Group based on WER . . . . .	55



# List of Figures

4.1	Gender Analysis in Irish Dataset . . . . .	23
4.2	Age Group Analysis in Irish Dataset . . . . .	24
4.3	Dialect Analysis in Irish Dataset . . . . .	25
4.4	Age Group by Dialect Analysis in Irish Dataset . . . . .	26
4.5	Dialect by Gender Analysis in Irish Dataset . . . . .	27
4.6	Age by Gender Analysis in Irish Dataset . . . . .	28
4.7	Duration of Audio Clips by Gender in Irish Dataset . . . . .	29
4.8	Duration of Audio Clips by Dialect . . . . .	30
4.9	Duration of Audio Clips by Age Group . . . . .	30
4.10	Sample audio signal . . . . .	31
4.11	Sample frequency distribution . . . . .	31
4.12	Frequency of Part-of-Speech Tags in Irish Dataset . . . . .	37
4.13	Irish Dataset Word Distribution (Top 50) . . . . .	38
5.1	Comparison of Model Performance . . . . .	44
5.2	Relationship between Training Samples and Performance Metrics . . . . .	46
5.3	Normalized POS tagging Vector Comparison . . . . .	48
5.4	Normalized Word Distribution Vector Comparison (Top 50 words from the Irish dataset shown) . . . . .	49
5.5	WER by Group Portuguese Dataset . . . . .	56
5.6	WER by Group German Model . . . . .	56
5.7	WER by Group Persian Model . . . . .	57

5.8	WER by Group French Model . . . . .	57
5.9	WER by Group English Model . . . . .	58
5.10	WER by Group Dutch Model . . . . .	58
5.11	WER by Group Arabic Model . . . . .	59
5.12	WER by Group Irish Model . . . . .	59
6.1	Gender Distribution Portuguese Dataset . . . . .	64
6.2	Gender Distribution Persian Dataset . . . . .	64
6.3	Gender Distribution Dutch Dataset . . . . .	65
6.4	Age Distribution Portuguese Dataset . . . . .	65
6.5	Age Distribution French Dataset . . . . .	66
6.6	Age Distribution Dutch Dataset . . . . .	66
6.7	Age Distribution Arabic Dataset . . . . .	67
6.8	Age Distribution Persian Dataset . . . . .	67
6.9	Gender Distribution German Dataset . . . . .	68
6.10	Gender Distribution French Dataset . . . . .	68
6.11	Age Distribution German Dataset . . . . .	69
6.12	Gender Distribution Arabic Dataset . . . . .	69
6.13	Gender Distribution English Dataset . . . . .	70
6.14	Age Distribution English . . . . .	70

# Chapter 1

## Introduction

This dissertation aims to look at how transfer learning can be used to improve the accuracy and performance of speech recognition on low resource languages. I have used Irish as the example for this project since it is a minority language that I know, but the goal is that the findings will be transferable to show how this work could be utilised for any language that is considered low resource. I am investigating the necessary steps to build a language model on a low resource language with limited data availability, the resulting quality of the model that can be obtained and how the speech recognition capabilities of the model can be improved through determining the optimal language to pre-train the fine-tuned model on.

### 1.1 Motivation

Low resource languages are often disadvantaged because they often do not have as many technological resources allocated to them as other, more widely spoken languages. This is due to them usually having less data available for research, a lack of funding for these languages and a smaller cohort of speakers. This is true for the Irish language. With over 39% of the population being able to speak Irish, we still have “relatively little progress in Automatic Speech Recognition(ASR)” Lynn (2023). This lack of fundamental technologies, specifically language technologies, has contributed to the process of language

extinction for Irish and has created a digital divide, resulting in speakers of the language reverting to using English, with only 1.5% of the population using the language outside of the education system Lynn (2023). These problems are seen across many low resource languages. to tackle these issues and bring these languages the technological advances they need, research is required. Resources, data and expertise in the language are all necessary to work towards a solution.

Imperial and colonial governments have been a major factor in the decline and extinction of many languages Chiblow and Meighan (2022), the Irish language being one such example. Irish was the main language spoken by the people of Ireland up until the 17th century Murtagh (2003). The Cromwellian plantations, planted English into the country and by the end of the 18th century, English had become the sole language of the government and public institutions Murtagh (2003). It was necessary for people to speak English in order to secure well paying jobs. Therefore, Irish was mainly spoken in the poorer areas of the country. With the famine of 1845 hitting these poorer communities the most, the language suffered. From the 1851 census, "approximately 45% of the population had spoken Irish during the last quarter of the eighteenth century" Murtagh (2003), this declined to 19.2% in 1891 and kept declining until the government started programs to revive the language and made it compulsory for schools to teach Irish Murtagh (2003). It seems the Irish language has not recovered in popularity since, even with the revival campaign. This is partly due to it lacking the modern technologies that other languages reap the benefits of. We need to invest time in technologies to assist it, specifically speech recognition technologies.

The effects of Irish not having the technological resources that other languages do were studied and it was found that there are challenges for learners of Irish, the most concerning of which being that learners do not have the opportunity to interact with native speakers of the language, partly due to teachers themselves learning Irish as a second language and due to the lack of native Irish speakers overall Chiaráin et al. (2022). Professionals have also been known to recommend that neurodivergent children not attend Irish language speaking school which is an old-fashioned view, not based on research Barnes et al. (2022).

This is a perfect example of how a lack of the right resources and assistive technologies for a language can lead to the exclusion of minority groups.

I would like to investigate how we can use pre-trained models to boost the performance of Irish ASR systems since it has such limited data availability.

## **1.2 Objectives**

The aim of this research is to comprehensively investigate the challenges faced by low resource languages, using the example of Irish in the context of ASR systems. First, a thorough analysis of the Irish dataset is needed to identify potential sources of bias, including gender bias and discriminatory language. This step is vital to ensure the development of fair and inclusive ASR models as it allows us to ascertain any bias that might exist therein. Next, gloss analysis techniques as a means of selecting the optimal source language dataset to use for pretraining as a means enhancing the recognition capabilities of low-resource languages like Irish will be investigated. Lastly, the efficacy of transfer learning to improve the speech recognition performance of Irish will be evaluated, thereby providing a technique to help advance the effectiveness of ASR systems for underrepresented language communities. With these objectives, I aim to contribute to the broader goal of promoting inclusivity and accessibility in ASR technology while addressing the specific challenges faced by low-resource languages like Irish.

With these objectives I would like to answer the following research questions:

- What are the challenges Irish faces as a low resource language?
- What bias exists in open source Irish datasets that are available to be used to train ASR systems?
- Can using gloss analysis help to identify the optimal pre-training language to use for fine-tuning Irish language models?

- What boost does transfer learning give to the speech recognition capabilities of Irish?

# Chapter 2

## State of the Art

This section is split in two. First I will define some topics in the Background section 2.1 that will appear throughout the dissertation. These definitions will be based on other researchers' work in areas related to the subject of this dissertation. In the second part of this section 2.2, I will give an overview of research done in the fields of machine learning/transfer learning, low-resource languages and bias in automatic speech recognition systems as these are the fields my work contributes to. I will also discuss important and noteworthy developments in those fields that are relevant to my project.

### 2.1 Background

This background section defines key topics in my report such as low-resource languages, transfer learning, and the Wav2Vec language model and its variations.

#### 2.1.1 Low-Resource Languages

Magueresse et al. (2020) defines low-resource languages as those that are “less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density”. They are languages typically spoken by minority groups, but most importantly for research purposes, they often lack the resources to train effective speech recognition models and perform natural language processing tasks. This means that as speech technology has

evolved throughout the years, technologies supporting these languages specifically have not been evolving with it. As described by Lonergan et al. , the result of this is an “unbalanced linguistic landscape” (Lonergan et al., 2023a). This has led to a scenario where we have speech technology that performs incredibly well for some languages, such as home voice assistants like Alexa (Amazon.com, Inc., 2024) and Google Home (Google LLC, 2024) while performing much worse for others. Much of the challenges faced when developing models for these languages stems from a lack of sufficient speech corpora, limiting the number of communities that can benefit from these technological advancements.

### 2.1.2 Transfer Learning

Transfer learning is a technique where a model is trained on one task and then adapted or *fine-tuned* to perform a different, but related, task (Torrey and Shavlik, 2010). Ideally, this means that the knowledge gained by a model from one task can be used to improve performance on a similar or related task. In machine learning, a model can be pre-trained on a certain language and then fine-tuned on another, with the knowledge gained from pre-training helping to improve the performance on the second language.

### 2.1.3 Wav2Vec

Wav2Vec (Schneider et al., 2019) was developed in 2019 and demonstrated the power of using self-supervised models for pre-training by learning representations from audio of speech alone, meaning this could be done on unlabelled data. The authors argue that this process is a lot more similar to how humans learn languages (Baevski et al., 2020). Prior to this development, language models were typically trained on large amounts of labelled data. This new Wav2Vec model allows for the development of models for speech recognition tasks using large amounts of unlabelled data to improve supervised models (Schneider et al., 2019). This model works by using a convolutional network to learn different aspects of the audio by extracting features. This is a “5 layer network with kernel sizes (10, 8, 4, 4, 4) and strides (5, 4, 2, 2, 2)” (Schneider et al., 2019). The context



network, consisting of “nine layers with kernel size three and stride one” (Schneider et al., 2019), takes these extracted features and combines them to gain more understanding of how this speech is formed, for example, the dependencies between words and the structure of the speech (Schneider et al., 2019). A contrastive loss function is used during pre-training to teach the model how to learn representations such that true samples are distinguishable from non-speech, known as *distractor* samples. This loss function which calculates the similarity between representations of true samples and the representations of distractor samples, encourages the model to learn representations which are closer to that of the true samples (Schneider et al., 2019). This model was able to achieve a low Word Error Rate (WER) of 2.43% using the Wall Street Journal test set (Garofolo et al., 1993) with a lot less labelled training data than was available to other models (Schneider et al., 2019).

## **Wav2Vec 2.0**

In 2020, this model was extended by Baevski et al. (2020) to make Wav2Vec 2.0 (Baevski et al., 2020). This paper showed the benefits from fine-tuning a model that learned its speech representation for audio without transcribed text. This is particularly useful for low-resource languages where labelled data tends to be scarce. Most of the time when these languages do have data available, it is unlabelled (Ranathunga et al., 2023). This was done by using a multi-layer convolutional feature encoder which extracts patterns and important parts of the audio. The Transformer-based context network helps to pick up on different dependencies between the words to encode context more effectively. This process is similar to what was done for the 2019 model but the techniques differ slightly. The audio signal is split up so the model can learn smaller, more manageable speech units and recognise these smaller sections of speech to make better predictions. A contrastive loss function is used during pre-training where some parts of the audio are masked and the model must predict what should be there – essentially filling in the gaps in the masked audio. The function maximises the similarity between the surrounding context and the prediction, and minimises the similarity with distractors, essentially picking the appropri-

ate representation (Baevski et al., 2020). Once the pre-training is complete, Baevski went on to fine-tune the model by “adding a randomly initialised linear projection layer onto the model” (Baevski et al., 2020) and defining the vocabulary of the speech is it to predict. Optimisation is done while fine-tuning by using a Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006) which helps train the model by transcribing the input speech into the corresponding sequence of characters that are defined in the vocabulary. It does this by calculating the probability of a character being the one that was input (Baevski et al., 2020). This approach to fine-tuning the pre-trained, self-supervised model was deemed very effective and was able to reduce the WER of the 2019 Wav2Vec model by about a third. (Baevski et al., 2020)

### **Wav2Vec 2.0 XLSR-53**

The Wav2Vec 2.0 model was extended by Conneau et al. (2020) to introduce a Cross-Lingual Speech Representation (XLSR) model by learning speech representations from multiple languages so that the model would generalise well across different languages. As mentioned in (Conneau et al., 2020), this model has the same architecture as described in 2.1.3 for the Wav2Vec 2.0 model. This means it is made up of a multi-layer convolution feature encoder followed by a transformer network. It is pre-trained on 53 languages and then fine-tuned on labelled data, again following the same procedure described in the Wav2Vec 2.0 model 2.1.3. The results found in (Conneau et al., 2020) show that this model significantly outperforms monolingual models, even when trained on the same amount of data, and the model achieves great results on low-resource languages. The aspect that is particularly interesting for my thesis is that better performance was achieved for low-resource languages when models were pre-trained on data from languages that are similar to the target language. The experiment that was done in this paper took 5 hours of Italian pre-trained with 50 hours of from other languages such as Spanish, German, English, Russian, Kabyle and Chinese. All models were then fine-tuned on 1 hour of Italian and the model that was pre-trained with Spanish had the best performance. Another interesting observation made in the same paper (Conneau et al., 2020) was that that multilingual

models seem to outperform monolingual models for low-resource languages due to positive transfer and being able to learn speech representation from multiple languages, but they perform worse for high-resource languages because the amount of data used in pre-training must be shared across languages.

## **2.2 Related Work**

A lot of research has been done regarding low-resource languages and the challenges that face them. From work that has been done at Trinity College Dublin and at other institutions it is clear that the classification of Irish as “definitely endangered” (Chiaráin et al., 2022) is accurate. I have broken down the difficulties that arise with these languages into different themes that I will be discussing in this section. These are themes relating to challenges with dialects, lack of data, bias in the data, problems with generalising across dialects, assumptions about the data, and lastly, the challenges in minority speech patterns. From the work done in this area it is clear that these topics apply to Irish as well.

### **2.2.1 Bias in Automatic Speech Recognition (ASR)**

A critical part of this research was to investigate the bias that exists in language models. It is well known that deficits in the data used to train models lead models to overlook certain groups in society (Bender et al., 2021). Bender et al. (2021) investigated this and also highlighted the fact that language models focus more on being able to handle tasks involving generating text that fit the grammatical rules required of it, rather than being able to do what a human does and understand the meaning of the words. The paper suggests that the problems begin with the data the model is being trained on, and that including bias in the data can lead to discrimination in the model’s results. An example of this can be seen in the investigation done by Liu et al. (2023) into discrimination against minorities, in this case women in the maritime industry. Liu et al. (2023) emphasised the importance of moving towards gender-inclusive language in the training

data to avoid discrimination and promote equality. Specifically focusing on the use of “seaman” versus “seafarer” in maritime speech. The paper notes that despite gender-inclusive pronouns being introduced in the 1980s, the biased pronoun “he” can still be found referring to a seafarer in maritime documentation. From the results they also saw that phrases such as “female seafarers” and “woman/women seafarers” occur more often than “seawoman”/“seawomen”. It seemed from their study of the industry “feminist language policies and guidelines in this field” (Liu et al., 2023) had more of an impact on the frequency of the word “seafarer” being used as opposed to more women actually working in these jobs and that changing the pronoun/word they use. This shows that once a gender-biased term exists in a certain industry, even if the culture of that industry changes, policies and guidelines are needed to change the terminology used. People working in the industry do not adapt organically, given the male-dominated nature of the environment, leading to the perpetuation of stereotypes.

When it comes to ASR systems and other language models, it is often the marginalised communities and minorities that bear the brunt of this discrimination. The discrimination takes various forms, from the BERT (Devlin et al., 2018) language model associating references to people with disabilities with more negative sentiments, to “gun violence, homelessness, and drug addiction” being overrepresented in texts related to mental illness Hutchinson et al. (2020).

Not only are these systems and language models biased in their outputs, but the environmental and financial cost of these models harm certain groups more than others. One example found that the training process of a certain Transformer model emits 284t of CO<sub>2</sub> (Bender et al., 2021). With the climate catastrophes around the world in the last few years such as wildfires in Australia and monsoons in Indian, the prioritisation of making these models more energy-efficient is long overdue (Bender et al., 2021). As is usually the case, it seems these negative outcomes are having the largest effect on groups outside of the Anglosphere, yet these are the groups seeing the least benefit since “over 90% of the world’s languages, used by more than a billion people, currently have little to no support in terms of language technology” (Bender et al., 2021).

There is bias that exists across all types of models, but what remains quite consistent is that it is extremely prevalent against low-resource languages, as there is usually a problem with gathering speech corpora that have enough variety to prevent discrimination against linguistic minorities who are not represented in the dataset. This was demonstrated in work done by Lonergan et al. (2023a) where dialects were removed and added to the corpora to see the effect. This experiment was done because of the challenges Irish language ASR systems face due to the lack of a “standard” for spoken Irish. The experimental procedure involved removing dialects from and adding them to the corpora to see the effect this has on the performance for each dialect. Removing dialects from the corpora, had a very negative effect on the performance of speakers with Ulster (UL) and Munster (MU) dialects, with the negative effect on speakers of the Connacht (CO) dialect being smaller in comparison. There was asymmetry observed between the performance of CO and MU, with MU being greatly impacted by the removal of MU speakers from the corpus compared to CO speakers, and CO’s performance not being affected much by the removal of either. Adding data for each dialect improved the performance for each dialect (CO having equal positive performance no matter what data was added but MU and UL experiencing a greater performance when their own dialect was added) (Lonergan et al., 2023a). This proved that a lack of representation in the pre-training data will lead to negative performance outcomes for the dialects which are underrepresented. Additionally, the lack of effect on the CO dialect contrasts with the argument that blindly adding more data for a better performance may not result in the improvement expected.

### 2.2.2 Specific Challenges For Low-Resource Languages

This section describes literature that address low-resource language-specific problems. One of the main problems comes about because, “in the ‘major’ widely spoken languages, technologies were developed for a standard variety” (Lonergan et al., 2022). This means that these technologies do not apply to Irish because there is not a a single spoken “standard dialect”<sup>1</sup>. It is very typical that low-resource languages don’t have a single spoken

---

<sup>1</sup>Though a written standard, *An Caighdeán Oifigiúil* (Tithe an Oireachtais, 2017) does exist.

standard. Approaches have been taken to try and deal with this challenge. ÉIST, an automatic speech recognition system for Irish has been developed, as explained by Lonergan et al. (2022), to deal with the variation in Irish dialects. They took two approaches to address these variations, the first being to create a pronunciation guide called the Trans-dialect set of rules so that the model would understand how different words and letters sound in Irish. To deal with the differences in pronunciation between the dialects, they added dialect-specific rules to this pronunciation guide. The second approach involved combining all these dialect-specific rules into one large pronunciation guide called Multi-dialect. The results found that the Multi-dialect rules and the Trans-dialect rules had a similar level of performance while occasionally the Trans-dialect rules performed slightly better. Lonergan et al. (2022) However we know that this effort is not put into larger language models to account for various dialects in a given language. Approaches involving transfer learning by Holmes et al. (2023) were used for the low-resource Irish Sign Language (ISL). Sign language datasets have a history of being biased with many having as few as six signers, showing how extremely low-resource they are. Holmes et al. (2023), improved the performance of the recognition capabilities of an ISL model by looking at how close the origins of other sign languages were to the origins of the ISL to find the optimal dataset to pre-train on. Following the process of pre-training on each dataset and fine-tuning on the Irish dataset, evaluating performance, and analysing vocabulary and lexical structure, they tried to find a dataset that would provide them “with the most attributes that overlap with our target dataset” (Holmes et al., 2023). Gloss analysis was recommended as a way to gain insights into vocabulary or words used in the dataset by considering the distribution of glosses, lemmas, and Part of Speech (PoS) tagging. The paper concluded by saying that finding the similarities between distributions is the best way to see if the dataset is best for pre-training, this is done with the cosine similarity technique. A performance boost is found by using pre-training and fine-tuning (Holmes et al., 2023).

Other challenges that low-resource languages face are that they are at the threat of extinction with it being said that “half of the presently living languages will become

extinct in the course of this century” (Adda et al., 2016). This is mainly due to a lack of funding for research on these languages. Without the resources in place to give these languages the technology they need to keep up with other more widely spoken languages, they will never flourish.

### **2.2.3 Challenges of Dialects**

Aside from the problems specific to Irish as a low-resource language, the main issue that researchers who investigated language models for Irish ran into was, that of dialects, as these models “must contend with a high degree of variability with limited corpora” (Loneragan et al., 2023b). The problems that can occur if the challenges with dialects are not properly addressed are disastrous, and they can occur for any language, not just Irish. It is even more prevalent now with smart voice assistants having a bigger impact on society since the pandemic with more people relying on them. For example, African-American users of these assistants have reported having to change their accent to make these systems better understand them. Additionally, we also see in an investigation done by Harwell (2018) for The Washington Post which found that Amazon Alexa devices produce 30% more inaccuracies for non-native accents. Bad performance across different dialects for English was confirmed again by work done by Koenecke et al. (2021), when it was found that in ASR systems such as those from Amazon, Apple, Google, Microsoft, and IBM had a WER 0.35 for Black speakers compared to 0.19 for White speakers when transcribing interviews from both groups of speakers. Again the bias against the African-American Vernacular English was found in DeepSpeech and in Google Cloud Speech with their models not being able to properly understand speech and the context of a sentence when the habitual “be” is used in this dialect (Martin and Tang, 2020). This further proves the problems these systems face when dialects vary from what the training data taught the systems to be the standard.

Without representation for all dialects of a language in these technologies, we see problems emerging such as those mentioned before by Chiaráin et al. (2022): that Irish language learners do not have many opportunities to hear the dialects of native speakers

of the languages, resulting in them finding it challenging to have conversations with them when they interact. People with specific dialects are forced to neutralise them in order to be understood by other speakers or ASR systems. Lonergan et al. (2023b) performed experiments on different models to be able to identify the different Irish dialects and found that an ECAPA-TDNN model, used for speech classification, trained from scratch on Irish and an ECAPA-TDNN trained on 107 languages and then fine-tuned on Irish both performed better than Facebook’s Wav2Vec XLSR model that was trained on 128 languages similar to the one described above 2.1.3. This shows how these large language models are not built to cater towards different dialects of a language, especially when it comes to low-resource languages.

## **2.2.4 Challenges of Minority Speech Patterns**

As well as dialect problems there are also issues in ASR systems with the speech patterns of minority groups, specifically for disabled people who have any kind of atypical speech pattern (Ngueajio and Washington, 2022). The technology is actually extremely biased against them with The Deaf or Hard of Hearing community experiencing a WER of 78% in ASR systems compared to a WER of 18% for hearing speech (Glasser et al., 2017). Glasser et al. (2017) gives many examples on how these technologies do not cater towards minority groups. It was found that the ASR systems did not work well in group environments and work better when used with speakers who have American accents. The study went on to say that if the systems were not being used in the ideal situation, (i.e. a controlled, lab like setting) then they were very ineffective and often not useful in real-world settings.

## **2.2.5 Assumptions About Our Data**

Considering the above literature on language models and the data that is used to train them, it is safe to say that a lot of assumptions are made about the information we are feeding these models, and that this can result in behaviour that can be very damaging



by promoting stereotypes and discriminating against certain groups. In research done by Fessler (2017), this problem was evidenced by the responses the technology gave when sexual comments were made towards them. It found that the technology did not actively discourage these advances and instead replied with encouraging replies such as “I’d blush if I could” (Fessler, 2017). This implicitly endorses the belief that sexual advances in inappropriate situations are acceptable. The responses also differed depending on whether a man or woman asked the question, with sexual advances from men warranting responses such as that described above, compared to responses such as “That’s not nice” for women (Fessler, 2017). This kind of behaviour is not a hard coded response, but rather patterns that have been learnt by the models due to the underlying bias in the data. This leads to the question: how should we go about building models and providing them with training data that won’t cause the issues we have spoken about?

An assumption that one would make is that you would create a balanced dataset, with equal representation across all groups of people. While there may alleviate some of the issues, it is not a complete solution. As stated above in Lonergan et al. (2023a) when experimenting with the removal of dialects from corpora and confirmed in Lonergan et al. (2023b) and Mengesha et al. (2021), it is not enough to only have a balanced corpora. In the experiments of Lonergan et al. (2023b), it was found that “balanced training corpora give rise to unequal dialect performance, with performance for the Ulster (UL) dialect being consistently worse than for the Connacht (CO) or Munster (MU) dialects.” Lonergan et al. (2023b) It seems that the UL dialect is considered more different and unique by these language models, making it easier to identify compared to CO and MU, which language models seem to have a hard time differentiating between. It would appear that there are subtle similarities between certain parts of our data that lead to differences in performance for different groups in our data. Therefore, we need to further investigate the contents of a dataset to make the models more equitable. We need to really understand the data and the language that is being used, as well as the characteristics of the speakers, in order to ensure an unbiased model.

Mengesha et al. (2021) also confirmed that more balanced and diverse training datasets

could improve the performance for minorities but that this alone is not sufficient. In particular, the authors suggests to involve community voices in solving issues with the recognition of their speech patterns. Mengesha et al. (2021) also mentioned the importance of having a way to correct the errors of the models with user feedback.

# Chapter 3

## Methodology

This section details the models that were used for the experiments, specifically the cross-lingual speech representation (XLSR) large-scale model described in 2.1.3 that I fine-tuned using various datasets. For the methodology of the project, I employed a two-fold approach to model development. I constructed a baseline model, which is a Wav2Vec 2.0 XLSR model fine-tuned on Irish. To explore the potential benefits of using pre-trained models, I adopted a transfer learning strategy. I fine-tuned models originally tuned for use on different languages, using the Irish dataset. The objective was to compare the performance of these fine-tuned models with the baseline model.

### 3.1 Data

The datasets used to train the large XLSR-53 model, which I used in my experiments, are a combination of Common Voice (a corpus with over two thousand hours of speech data in 38 languages) (Mozilla Corporation, 2021), BABEL (Gales et al., 2014) (a corpus of telephone conversation from many languages), and Multilingual LibriSpeech (a corpus made of speech from audiobooks across 8 languages) that results in a large corpus of 53 languages (Conneau et al., 2020). This Wav2Vec XLSR-53 model is trained on unlabelled data with the ability to be fine-tuned on labelled data and can be used to enhance different speech recognition tasks.

## 3.2 Fine-Tuned Models

In my experiments, I used XLSR-53 models that were pre-trained on 53 different languages from three different large datasets, and then fine-tuned on specific languages (Grosman, 2021). During the fine-tuning process the model is trained on labelled data. This fine-tuning process involves adding a linear layer on top of the pre-trained model to transform the representations learned during pre-training (the high-dimensional embeddings) into predictions, usually referred to as logits. These logits are made to align with the output vocabulary of the target language, Irish, allowing for the computation of probabilities across the possible outputs. A CTC (Graves et al., 2006) loss function, is used during this training which adjusts the models parameters to minimize the difference between the models predictions and the ground truth – in my case the audio transcriptions of the labelled data (Conneau et al., 2020). By doing this, the learned representation from the XLSR-53 model is fine-tuned to the labelled data. I chose to use models that had already been fine-tuned on specific languages using Common Voice, from a hugging face library (Grosman, 2021). They are fine-tuned facebook/wav2vec2-large-xlsr-53 models using the train and validation splits of Common Voice 6.1. Some languages such as Dutch are underrepresented in Common Voice with it only having 64 recorded hours in the 6.1 version dataset. If we compare this to the 837 recorded hours in the Common Voice German dataset or the 2,182 recorded hours of the English dataset, we notice a huge difference. Since it would be deemed unfair to try and compare these models when there is such a disparity in the amount of data available for different languages, where possible, extra audio clips were taken from CSS10, a collection of single speaker speech datasets for 10 languages, by Grosman (2021) to be added to the smaller datasets. Each of these datasets consists of audio files recorded by a single volunteer and their aligned text sourced from LibriVox (Park and Mulc, 2019). From the models provided by Grosman (2021), I used those that were already fine-tuned on Arabic, Dutch, French, German, Persian, Portuguese and English ? for these experiments. Only the models fine-tuned on Dutch and Arabic contain extra data from CSS10.

### 3.3 Fine-Tuning Models on Irish

To fine-tune the Irish baseline model referred to above, I used the following steps:

1. The Common Voice dataset version 15.0 (Mozilla Corporation, 2021) for Irish with their corresponding training, validation, and testing splits. This is labelled data as opposed to the unlabelled data that the self-supervised XLSR-53 model is trained on.
2. The preparation of the data involved removing special characters and unnecessary columns from the transcripts of the audio and making all the words lower case.
3. A vocabulary was then built using tokenisation. All the unique characters from the training and test set are collected. All these unique characters are mapped to a number and saved as a vocab-to-id mapping.
4. The pre-processing of the audio files involved ensuring that audio was the required sampling rate. In this case, all audio is sampled at 16kHz. The files are then padded.
5. The appropriate hyperparameters are chosen for this task by performing Bayesian hyperparameter optimization (Yang and Shami, 2020), which searches for the best hyperparameters of the model by finding the maximum or minimum of an objective function, in my case I chose that to be the Word Error Rate (WER). It was not important for me to fully optimise this model as the goal of this dissertation is to compare models rather than find the best possible model for Irish ASR. Realistically when training a model on a low-resource language, it is very challenging to achieve a near-perfect performance.

Once the fine-tuned Irish model is optimised (enough), I then have the defined hyperparameters that I will use when training all other models. The hyperparameters were kept the same throughout the creation of all the other models to see which gave the best performance boost.

To be create the other fine-tuned models, I followed the Irish dataset preparation and pre-processing steps 1–5 and then the following:

6. The pre-trained model is loaded in. In this case the “pre-trained” model is the model fine-tuned on a certain language created by Grosman (2021) that was pre-trained on facebook/wav2vec2-large-xlsr-53.
7. The model is fine-tuned using the same hyperparameters optimised for fine-tuning on the Irish dataset.
8. The feature extractor part of the model is frozen so that the weights of the feature extractor are not updated during training with only the later layers being fine-tuned. This means that the useful features and valuable information that the feature extractor learned during pre-training, is not modified during the fine-tuning process. This can also help to prevent overfitting.

All fine-tuned models for the experiments were created using the same steps, and the exact same hyperparameters. This only difference between them was that they used fine-tuned models from different languages in step 6 before fine-tuning on Irish.

# Chapter 4

## Experimental Setup

In order to properly analyse the model, I fine-tuned on different languages. To evaluate them effectively, we needed to have a good understanding of the data that they are being trained on, as discussed in 2.2.5. As mentioned in Barnes et al. (2022), when developing a system specifically designed for the Irish language, expertise is needed in areas such as Irish semantics, syntax and morphology in order to notice any errors in the data or results and to overcome them. As described by Barnes et al. (2022) “It is an inflected language, with a number of cases for nouns and adjectives as well as inflected prepositions and verbs”. This means that a system already in place for communicating through a language such as English will require significant alterations in order to adapt it to a language such as Irish. Specifically Barnes et al. (2022) provides the example of counting in English, where there is a unique set of ordinal numbers, compared to Irish where there are two sets of ordinal numbers: one for counting people and another for counting objects. Barnes et al. (2022) also notes that phrasal verbs in Irish change meaning depending on whether they are followed by a preposition (“ag éirí” versus “ag éirí le”). Barnes et al. (2022) also points to the manner in which Irish morphemes are often added to English words which may not have a current corresponding Irish translation, such as the verbal noun “ag zoomáil”, which uses the modern verb “to zoom” in its construction. Such elements, which are distinct to the Irish language, can only be noticed by someone who has knowledge of the language itself. As an Irish speaker and English speaker, I can perform analysis on the

Irish dataset to deduce patterns, errors, or strange sentences etc., however, when analysing the datasets from the other languages, I can only gather statistics and compare them to those gathered for Irish. I can notice patterns once the datasets have been translated to English but I can't be sure if these patterns came about because of the translation or that data itself.

## 4.1 Dataset

I did an in-depth analysis of each dataset that was used to fine-tuning the models to gain understanding of their characteristics and suitability for pre-training. This involved gathering statistics on the speakers, checking for bias in the datasets, and gloss analysis of the datasets to determine the most effective one to pre-train on based on similar vocabulary and lexical structure.

### 4.1.1 Common Voice

The data used to fine-tune models came from the Mozilla Common Voice datasets. Common Voice is a project by Mozilla that contains free datasets that can be used for speech recognition. It is also a crowd-sourced dataset, where people's voices can be submitted and are turned into MP3 files with a corresponding text file transcription of the audio. These volunteers can also help to validate the recordings of other. Datasets are split up into files titled train, test, validated, invalidated, reported, etc. A times file is included with the length of each audio file specified. For the analysis below and for fine-tuning any models on Irish, I used the train, test and validated files from the Irish Common Voice Corpus 15.0. These files contain the metadata for the audio with the headings listed in Table columns 4.1.

client_id	path	sentence	up_votes	down_votes	age	gender	accents	variant	locale	segment
-----------	------	----------	----------	------------	-----	--------	---------	---------	--------	---------

Table 4.1: Metadata Tags for Irish Dataset



## 4.2 Data Analysis of the Irish Common Voice Dataset

Below contains the analysis of the Irish dataset. It is important to note that a considerable portion of the audio files in the dataset lacked some associated metadata. For example in Figure 4.1, out of the 536 total audio clips in the training set, 205 audio files do not specify the gender of the speaker.

### Gender Analysis

Of the 536 clips in Figure 4.1, 294 clips contain a male voice. There are 37 clips containing female voice. 205 clips have no gender associated with them. That is a 14:1 ratio between men and women.

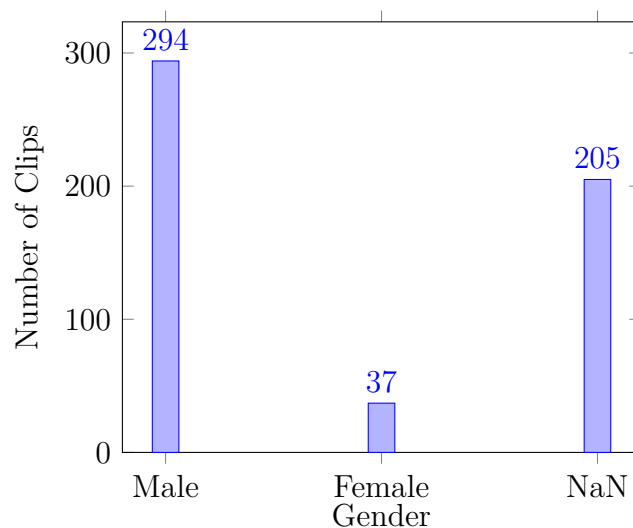


Figure 4.1: Gender Analysis in Irish Dataset

There are a lot more males recorded in the metadata than females, making the data initially seem quite biased. It could be the case here that women felt uncomfortable labelling themselves as they are cautious of the stereotypes associated with their gender, this will be touched on more in the Section 4.3. Nevertheless, around 38% of the metadata does not have a gender associated with it so this could account for the apparent bias in the dataset.

## Age Analysis

According to Figure 4.2, the age group with the largest representation are people in their thirties. The distribution is quite varied:

- Ages 30-39: 31.34%,
- Ages 20-29: 17.16%,
- Ages 50-59: 8.78%,
- Teens: 4.48%,
- Ages 40-49: 0.19%,
- No recorded age: 38.22%.

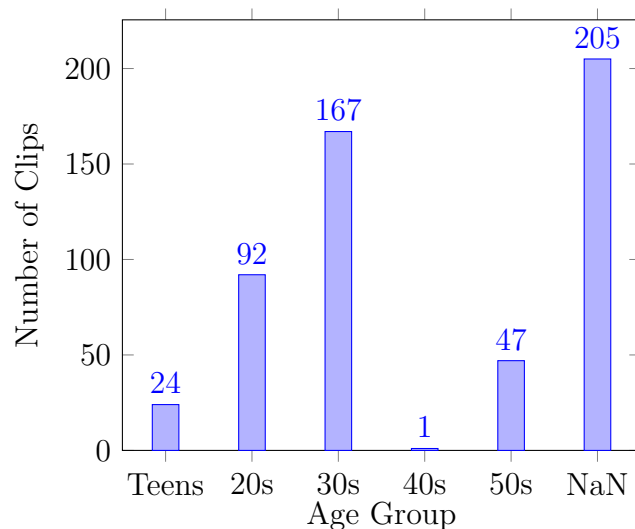


Figure 4.2: Age Group Analysis in Irish Dataset

We see a large range of ages represented in the dataset which is good but once again, much of the data does not have an age assigned to it. There might be some bias in the dataset since some (particularly older) age groups are not represented at all. We don't see anyone aged 60 or above represented and the youngest age group seen are teenagers.

## Dialect Analysis

All three Irish dialects are represented in the dataset as seen in Figure 4.3, with the following distribution:

- Gaeilge Chonnacht: 18.66%,
- Gaeilge Uladh: 13.26%,
- Gaeilge na Mumhan: 12.69%,
- Unlabelled: 55.42%.

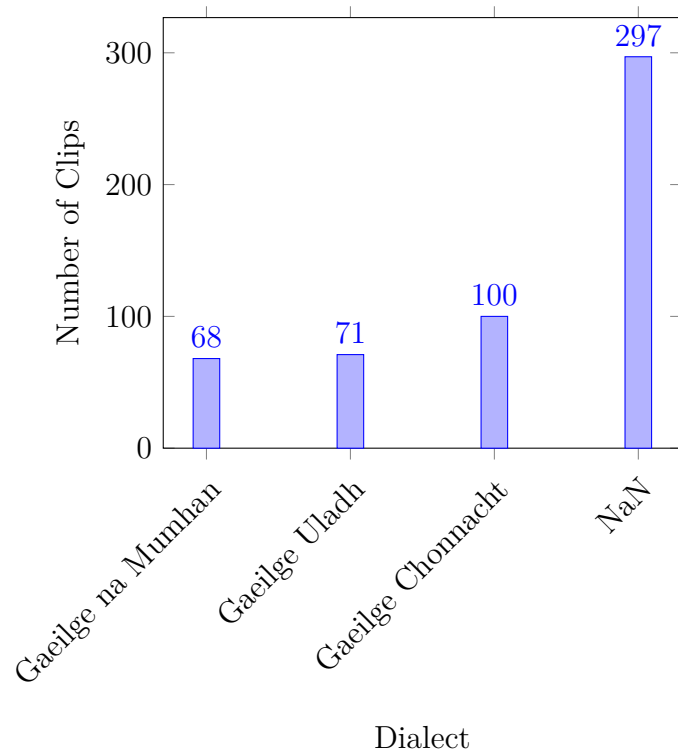


Figure 4.3: Dialect Analysis in Irish Dataset

Most of the dataset again contains unlabelled information which can indicate a potential problem with dialect classification. Of the data that is labelled, it is distributed reasonably evenly.

### Dialect by Age Analysis

In Figure 4.4, we can see the distribution of dialects by age groups. For the teens age group, 24 clips are associated with Gaeilge Chonnacht. For people in their twenties, 37 clips are associated with Gaeilge na Mumhan. For people in their thirties, 100 clips are associated with Gaeilge Chonnacht, and 30 are associated with Gaeilge na Mumhan. For those in their forties, only one clip is associated with Gaeilge na Mumhan and for people in their fifties, 47 clips are associated with Gaeilge Uladh.

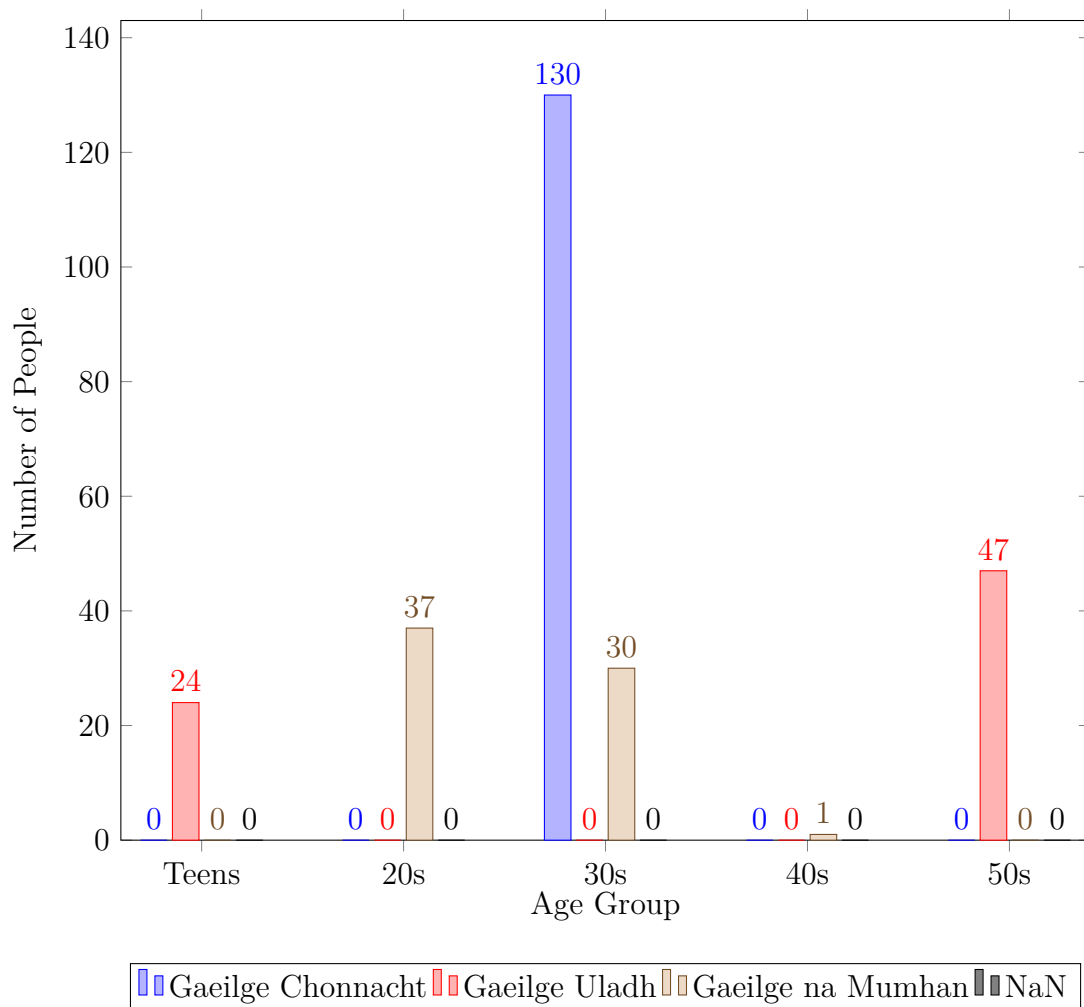


Figure 4.4: Age Group by Dialect Analysis in Irish Dataset

This means that for Gaeilge Chonnacht, clips from people in their thirties only are associated with this dialect. For Gaeilge Uladh, clips from people in their fifties and teens are associated with this dialect and for Gaeilge na Mumhan, clips from people in the

twenties, thirties, and forties are associated with this dialect. Gaeilge Uladh is distributed across two age groups while Gaeilge na Mumhan is seen across three. This distribution was interesting to note as it posed the question as to whether the concentration of certain dialects with specific age groups may reflect regional patterns or whether it is just a bias of the dataset.

### Dialect by Gender Analysis

All the clips with dialect metadata from the train.tsv file were recorded by males as seen in Figure 4.5.

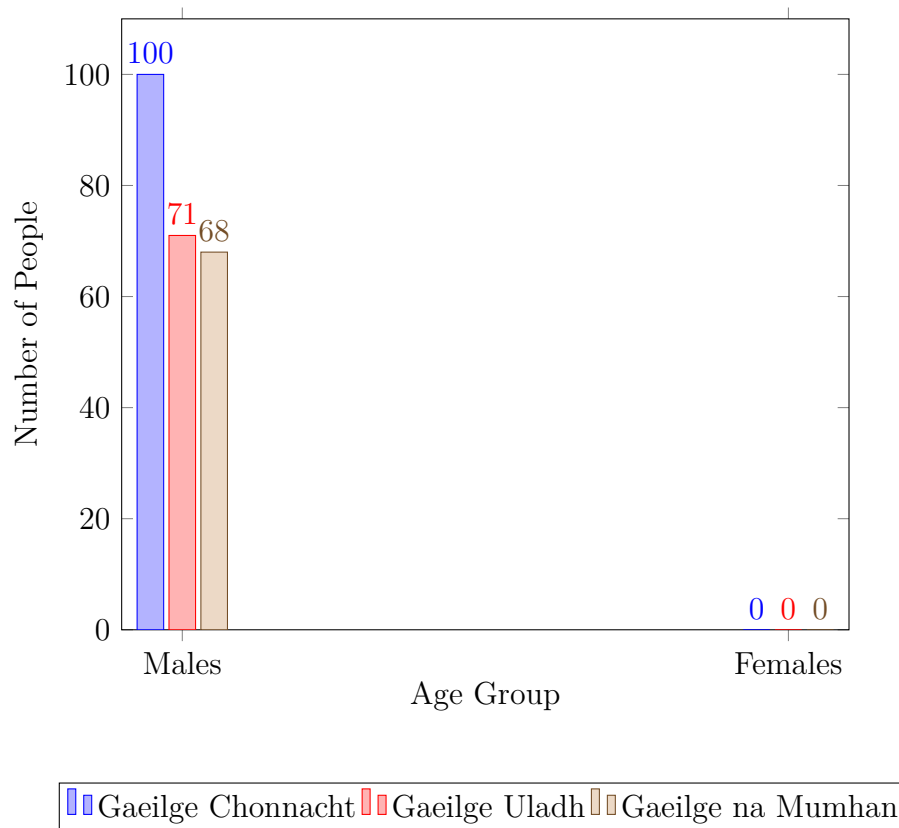


Figure 4.5: Dialect by Gender Analysis in Irish Dataset

This could lead to a lot of difficulty for the model when trying to understand females with different dialects. We know that we have female speakers in the dataset but their dialect are not labelled so the distribution is unknown.

### Age by Gender Analysis

We can see here that all clips recorded by people denoted as female in the metadata fall in the twenties age group. No other female age group disclosed their age.

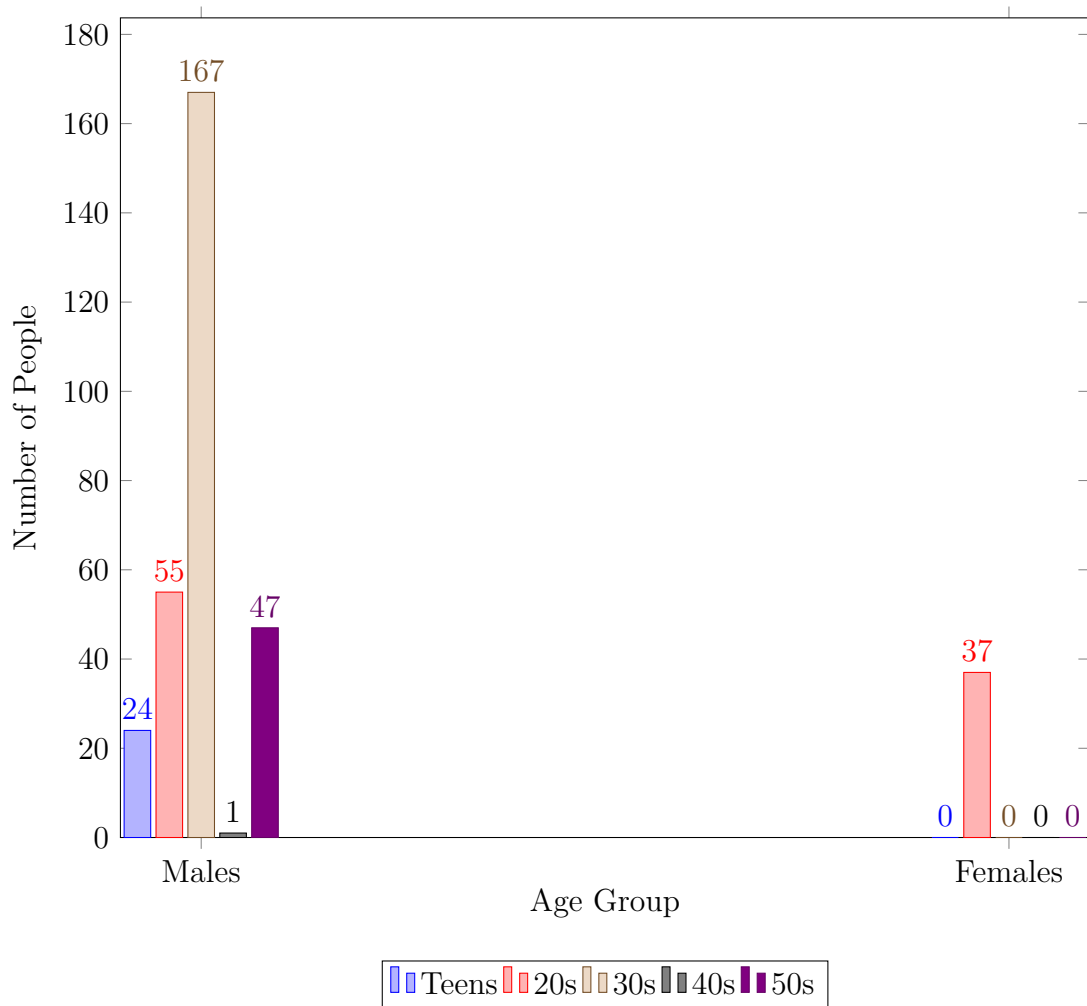


Figure 4.6: Age by Gender Analysis in Irish Dataset

Within the male age group, most of the clips are recorded by men in their thirties. There is quite a good distribution of ages among males with some bias towards men in their thirties and only a single clip being recorded by a man in their forties.

### Duration of audio clips by gender

We know that there is a 14:1 ratio between the amount of male recorded clips versus female recorded clips. When we look at the length of time of male audio clips we are using to train on compared to that of female audio clips in 4.7, we see that there is also a huge bias with a 54:7 ratio, showing that the model will be trained for significantly more time on male voices.

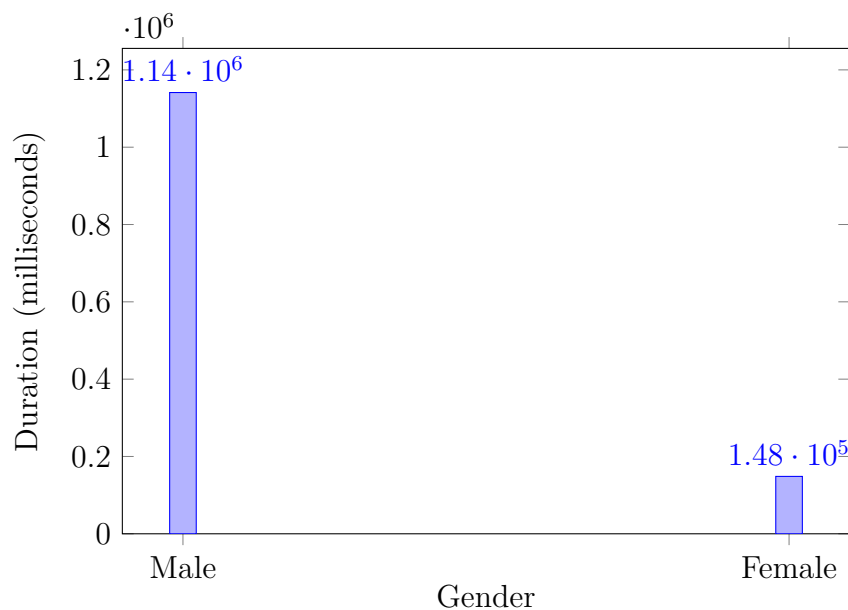


Figure 4.7: Duration of Audio Clips by Gender in Irish Dataset

### Duration of audio clips by dialect

From Figure 4.8, we can see there is a difference of 58044ms between the length of time of Connacht dialect clips compared to that of Ulster dialect clips. There is a bigger difference of 61704ms between the length of time of Ulster dialect clips compared to that of Munster dialect clips. We can see there is a significant difference between the length of audio clips we have for the Connacht dialect compared to that Munster dialect clips, 119748ms. The amount of audio clips per dialect seems to follow the same pattern as the amount of clips per dialect with Gaeilge Chonnacht coming out on top, followed by Gaeilge Uladh and lastly, Gaeilge na Mumhan.

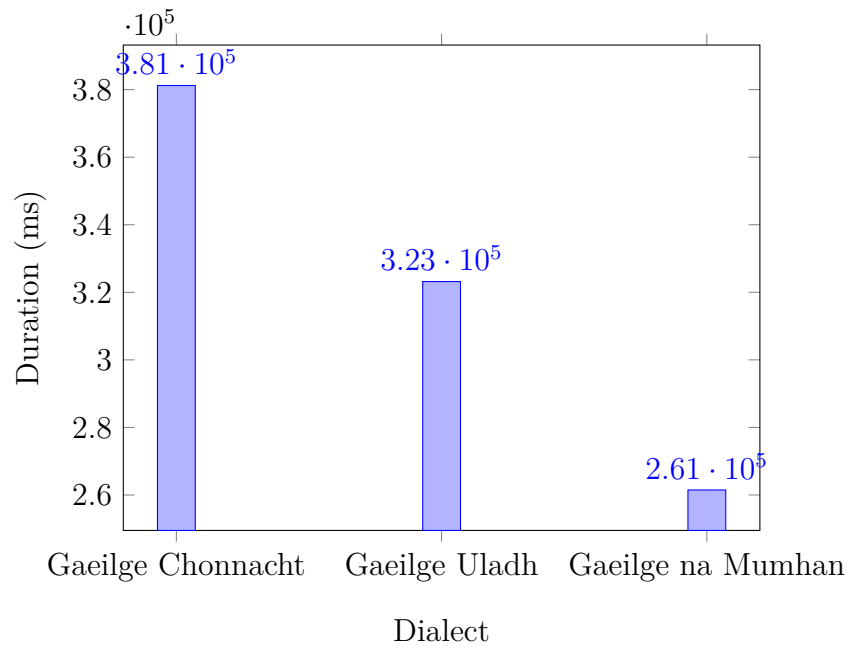


Figure 4.8: Duration of Audio Clips by Dialect

### Duration of audio clips by age group

For the length of audio for each age group in Figure 4.9, there is the same pattern here to what was found in the distribution of recordings by age group (the amount of audio clips for each age group) in Figure 4.2, with the thirties age group having the longest amount of audio associated with it, followed by twenties, fifties, teens and lastly forties.

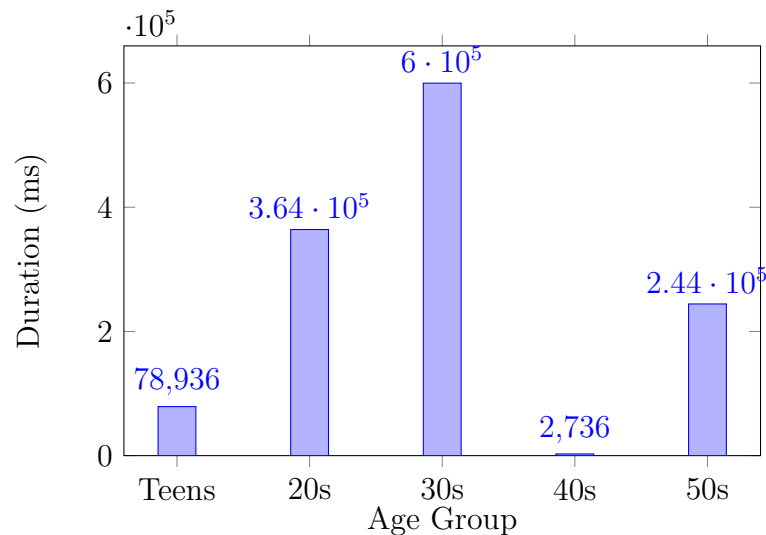


Figure 4.9: Duration of Audio Clips by Age Group



## Duration of audio clips, sampling rates, and frequency distribution

To ensure that all clips contained audio and that nothing was amiss with the clips, I checked the times.txt provided in the Irish common voice dataset which gave details of the clip durations. I noted the average and median duration of the clips (3853ms and 3576ms) respectively. I also saw that the minimum duration of a clip in train.tsv was 1224ms, meaning there are no clips of duration "0 seconds", which would indicate a clip without any sound.

The values seen for the sampling rate were all 48kHz which is common for audio recordings, except for 35 clips with a lower sampling rate of 32kHz. I inspected the audio for all of them to check that there was speech in the clips.

I checked all audio signals in the time domain using features such as in Figure 4.10, and all the time-frequency representations of the audio files, such as that shown in Figure 4.11, for any strange patterns. Any abnormalities such as sudden peaks, indications of too much background noise, and indications of no sound, were noted and these recordings were replayed, but no faulty audio clips were found.

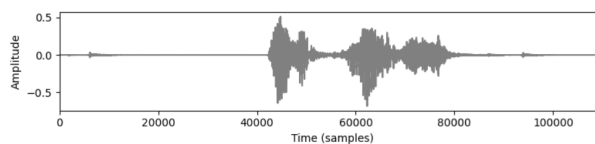


Figure 4.10: Sample audio signal

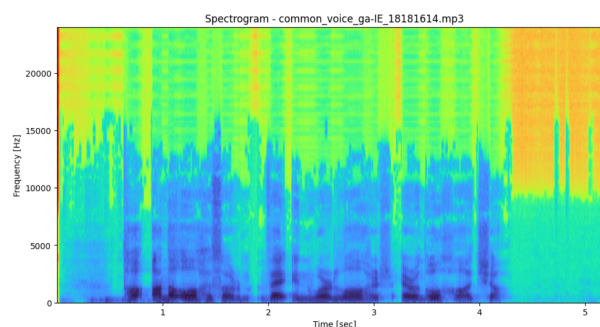


Figure 4.11: Sample frequency distribution

### 4.3 Results of Investigation into Data Analysis Results of the Irish Common Voice Dataset

From the inspection of the clips, I can confidently say that the audio quality in this dataset is sufficient to build models with. I thought it would be interesting to check if the statistics here represent a bias in the dataset or are indeed a true representation of the population of Irish speakers. For example, perhaps it is the case that there are a lot of Irish speakers in their thirties with a Connacht dialect, etc. I checked the CSO data on the Irish language (Central Statistics Office (Ireland), 2016). From the census it was shown that more females than males identified themselves as Irish speakers, 55% compared to 45%, therefore self-described males were over-represented in the dataset used in this section, see Figure 4.1, with the actual number of males and females being unknown due to a lack of complete metadata.

As regards the distribution of ages of Irish speakers, from Table 4.2, we can see that the distribution of ages in the dataset does not match that of the general population of Irish speakers.

Table 4.2: Age groups of the Irish Dataset vs Irish speaking population

Age Group	Common Voice Dataset (%)	Irish Population(%)
Teens	4.48	26
Twenties	17.16	11.5
Thirties	31.34	14.5
Forties	0.19	19.2
Fifties	8.78	9.7

If we look closer at the distribution of ages by gender in Table 4.3, we have a similar percentage of females in their twenties in the dataset as there are in the general population. We cannot see the distribution across the other age groups because this information was not documented in the dataset.

Looking closer at the distribution of males across the various age groups in 4.4, the distribution of males by age in the dataset is dissimilar to that in the general population.

Unfortunately, that was the only information I found useful to my analysis from the 2016 census. There was no detailed information regarding the dialects of Irish speakers in the census. Based on the observations of this subsection, we can say there is a strong indication of bias in the dataset.

Table 4.3: Female age groups of the Irish dataset Vs Irish speaking population

Age Group	Dataset Female (%)	Population Female (%)
Teens	NaN	8.3
Twenties	6.9	6.5
Thirties	NaN	8.5
Forties	NaN	7.1
Fifties	NaN	5.7

Table 4.4: Male age groups of the Irish dataset Vs Irish speaking population

Age Group	Dataset Male (%)	Population Male (%)
Teens	4.47	7.7
Twenties	10.3	5
Thirties	31.2	6
Forties	1.9	12.1
Fifties	8.8	4

As for the gender bias in the dataset, only 6.9% of women disclosed their gender in the Irish dataset 4.6. It is most likely that there are more female speakers in the dataset but they didn't label themselves as being female. There are many factors that can contribute to women not wanting to disclose their age but it is most likely due to the stereotypes and prejudice against women. A study of this was done by Quéniart and Charpentier (2012) which looked into how older women feel they are perceived as dependent and are defined as fragile once they reach a certain age. This is not necessarily how they feel themselves so they tend not to disclose their age so as not to be thought of in this way. Sexist and ageist comments can “naturally have an impact on women, especially on their self-esteem.” (Quéniart and Charpentier, 2012). I would say it is highly likely that the same thing happened here when women were choosing what data to label themselves

with. Only women in their twenties decided to disclose their age confirming what was said in Quéniart and Charpentier (2012) about women being hesitant to define themselves as old. Other women probably didn't provide labelled metadata with their audio for similar reasons.

## 4.4 Text Analysis of the Irish Common Voice Dataset

I did text analysis on the dataset to extract any insights, patterns and trends from textual data. By delving into the linguistic characteristics of the dataset, I aimed to uncover underlying patterns that could inform further analysis and model development. In order to compare this dataset with others, I translated the sentences into English, as I will do with the other datasets. This provided a standardised foundation for conducting text analysis. I translated the sentences for all datasets using the open source argostranslate<sup>1</sup>, a Python library for translating between different languages. I did notice that some translations for Irish are not fully correct. To give a few example, “Tá bean bheag ag an doras, agus tá bróga bána uirthi”, got translated to “The door has a small woman, and she has white shoes”. The translation seems to have misunderstood the the preposition “ag an” which in English would translate to “at”. In Irish the word “Tá”, meaning “is”, can be combined with the word “ag”, to indicate possession. This is where the mistranslation has come from in this case. Another example of an inaccuracy from the translation of argostranslate is, “Drochrud a ghabh thar an doras arsa Bairbre”, which was translated to “Drochrud caught over the door of Bairbre”. Here no translation was found for the word “Drochrud”, meaning “bad thing”. This seems to have lead to the mistranslation to the rest of the rest of the sentence since the translation didn't fully represent what was being communicated in the text.

Once the sentences from the training data were in English, I tokenised them using RegexpTokenizer<sup>2</sup> from the NLTK library, a class used to tokenise using regular expres-

---

<sup>1</sup><https://www.argosopentech.com/>

<sup>2</sup><https://www.nltk.org/api/nltk.tokenize.RegexpTokenizer.html>

sions, and `re`<sup>3</sup>, Python's built-in regular expression library to remove punctuation. Each sentence's punctuation was removed using this regular expression pattern.

```
[^\w\s]|\_
```

It will match any character that is not alphanumeric or whitespace, or an underscore. Once this was done the sentences were tokenised.

#### 4.4.1 Part-Of-Speech(POS) tagging

POS tagging is the process of the word in each sentence being assigned a grammatical label based on its syntactic role within the sentence. I perform this using NLTK's `pos_tag`<sup>4</sup> function on all of the word tokens, which results in an example like the following:

```
('A', 'DT'),  
( 'member', 'NN'),  
( 'of', 'IN'),  
( 'the', 'DT'),  
( 'Minister', 'NNP')
```

I iterate over these tuples, removing the first element (the word), and keep the tag. These tags are stored in a flatten array. I use the `Counter`<sup>5</sup> class from the `collections` module in Python to count the occurrences of each unique tag in the flattened array, resulting in a list of tuples where each tuple consists of a POS tag and its corresponding frequency count. The most and least common tags in the dataset are given in Tables 4.5 and 4.6 respectively.

I graphed these frequencies for further inspection into the dataset. As seen in Figure 4.12, the distribution of the frequencies here looks quite good with there being a lot of nouns, preposition, adjectives, etc in the dataset. We also have a wide variety of tags here which is good for training a model of different types of words. The frequency chart here looks quite common and doesn't seem abnormal.

---

<sup>3</sup><https://docs.python.org/3/library/re.html#module-re>

<sup>4</sup>[https://www.nltk.org/api/nltk.tag.pos\\_tag.html](https://www.nltk.org/api/nltk.tag.pos_tag.html)

<sup>5</sup><https://docs.python.org/3/library/collections.html#collections.Counter>

Table 4.5: Most Common Part-of-speech tags , meanings, and examples (In order of frequency)

<b>Tag</b>	<b>Meaning</b>	<b>Example</b>
NN	Noun, common, singular or mass	"world"
DT	Determiner	"the", "a"
IN	Preposition/subordinating conjunction	"of", "with"
NNP	Proper noun, singular	names/places
PRP	Personal pronoun	"she", "he"
JJ	Adjective	"big", "small"
VB	Verb, base form	"take", "talk"
NNS	Noun, plural	"shops"
VBP	Verb, non-3rd person singular present	"I talk"
VBZ	Verb, 3rd person singular present	"talks"
RB	Adverb	"slowly"
TO	To	"to do"
CC	Coordinating conjunction	"and", "or"

Table 4.6: Least Common Part-of-speech tags, meanings, and examples

<b>Tag</b>	<b>Meaning</b>	<b>Example</b>
UH	Interjection	"ah"
FW	Foreign Word	words from another language
NNPS	Proper Noun, plural	plurals of names/places
RBS	Adverb, superlative	"best"
JJR	Adjective, comparative	"bigger"
PDT	Predeterminer	"both", "all"
RBR	Adverb, comparative	"stronger"

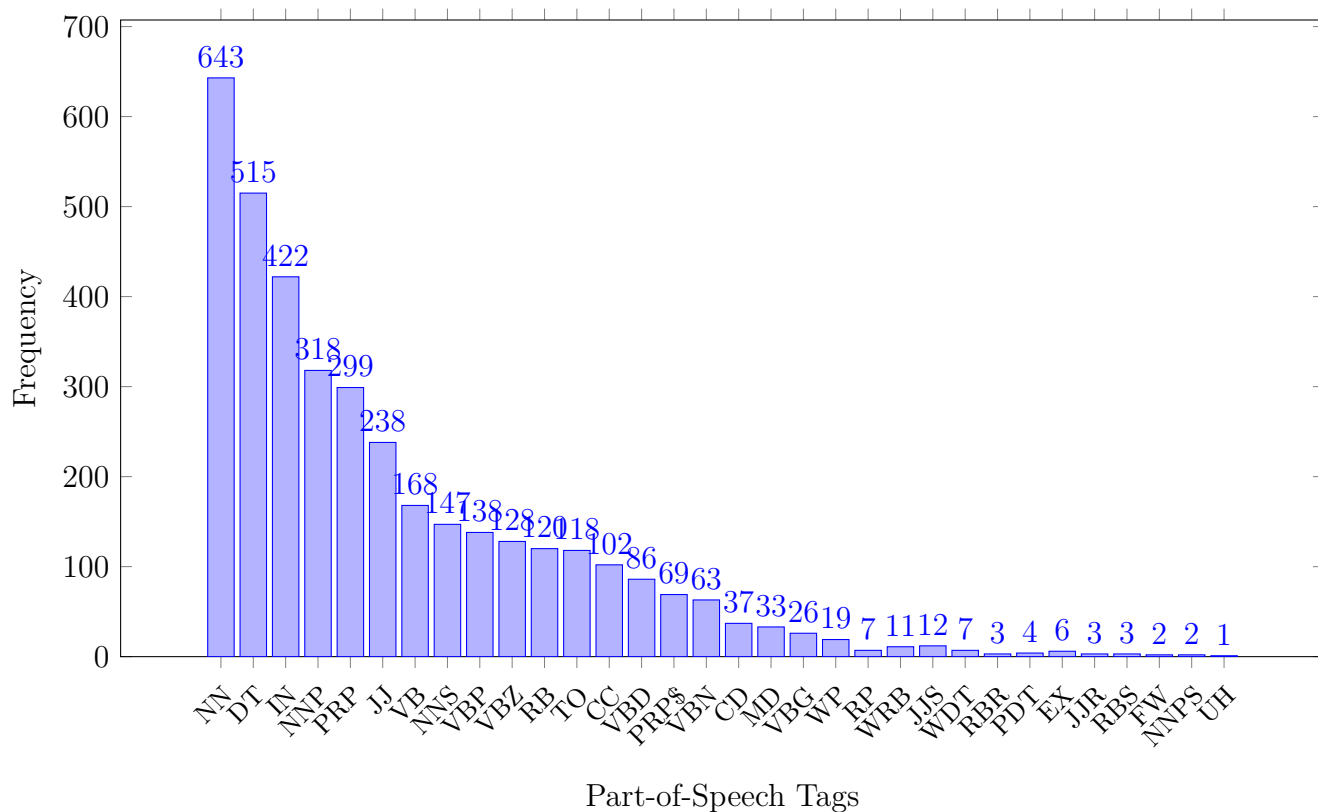


Figure 4.12: Frequency of Part-of-Speech Tags in Irish Dataset

## 4.4.2 Word Distribution

To look at the frequency of different words in the dataset I needed to first reduce the words to their base form using lemmatisation, meaning different inflected forms of a word would be treated the same. I use the `WordNetLemmatizer`<sup>6</sup> class from the NLTK library. Each word in each sentence was lemmatised. Some behaviour I noticed from lemmatising was that "as" gets mapped to the base form "a", this could increase the frequency of the word "a" and "was" is lemmatized to "wa", which is incorrect.

Once complete, the arrays of lemmatised words are flattened and using the same counter class as before to count the occurrences of each unique word. This results in an array of tuples such as the following:

```
('a', 154),
('member', 3),
```

<sup>6</sup><https://www.nltk.org/api/nltk.stem.wordnet.html?highlight=wordnetlemmatizer>

('of', 133),  
 ('the', 277),  
 ('minister', 4)

To properly visualise these frequencies I plotted them on different subplots.

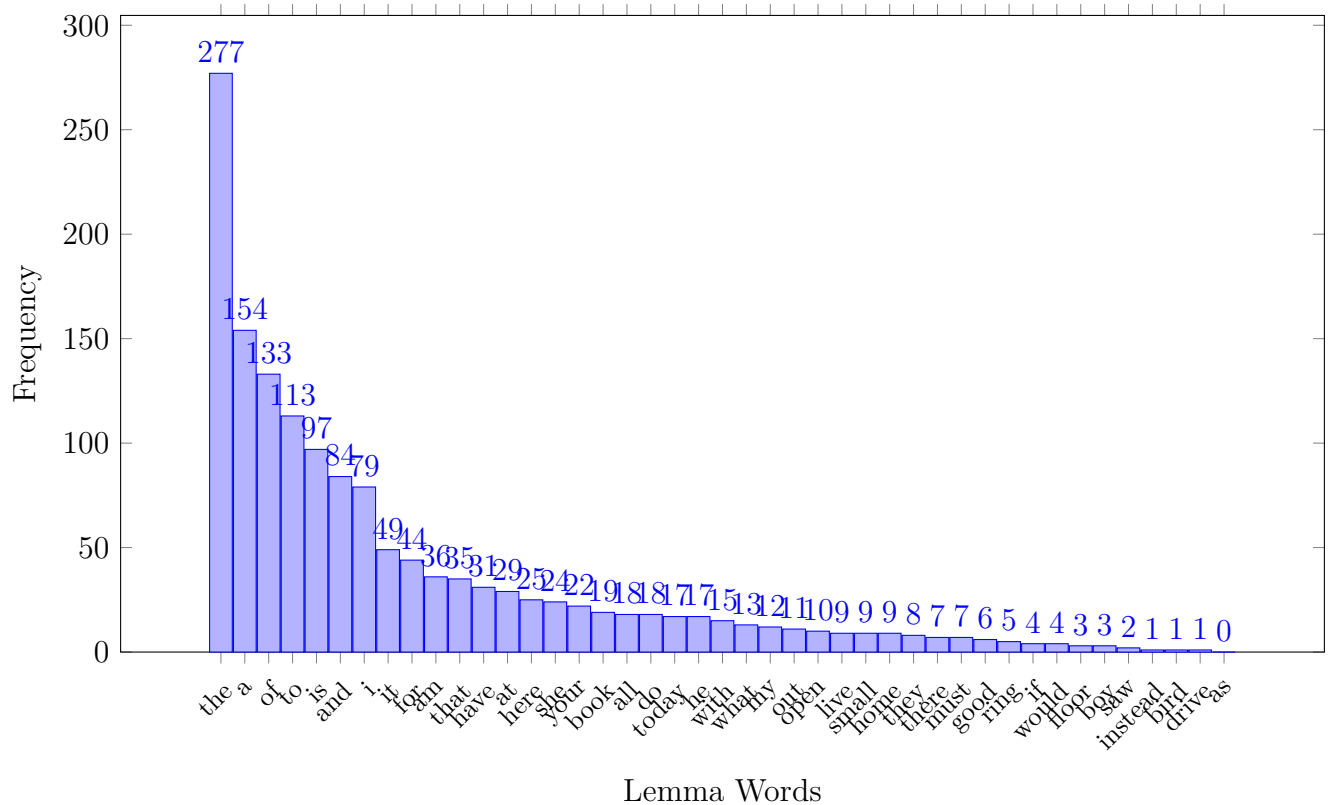


Figure 4.13: Irish Dataset Word Distribution (Top 50)

From Figure 4.13 we can see the highest frequency words include those you would expect to see in everyday sentences such as: “the”, “a”, “of”, “to”, “is”, “and”, “i”, “you”, “it”, “in” and “for”. Some other words that were common are: “book”, “irish”, “school”, “pleasure”, “home”, “community”, and “language”. The idea is this analysis gives the ability to compare gloss analysis across different datasets.



## 4.5 Results of Investigation into Text Analysis Results of the Irish Common Voice Dataset

Upon further analysis into the Irish dataset I noticed some themes in the vocabulary. I saw a strong political/historical theme in the dataset with references to words such as “prison”, “nationalist”, “republican”, “socialist”, “boycott”, “radical”, “political”, “history”, “saorstát” and “gael”. There were also some reference to religion, such as “god”, “saints”, “soul”, “death”, “pray”. These themes would be expected since, even the word for “Hello” in Irish ,references god and religion would be common because of the influence of Catholicism on the language. Words that relate to both these themes would be seen in common phrases in Irish. The theme of education is common in a lot of language datasets and was evident here as well, with the presence of words like “preschool”, “university”, “institute”, “school”, “academic” and “educational”.

I took a further look into the gender bias that exists in the dataset. I wanted to analyse how men and women were spoken about. I noticed that many names were referenced in the dataset, particularly male names, and I thought it would be interesting to further analyse this. There were 1033 lemmatised words in the Irish dataset, 42 of which were first names. Of the 42 names, 64.3% were male names and 35.7% were female names. This indicates a possible disparity in the ways men and women are referenced. There are 97 lemmatised words directly referencing men in the dataset between first names and personal pronouns, 38% on a first name basis. Of the 63 lemmatised words referencing women, again between first names and personal pronouns, 28% are on a first name basis. “She” was reference 24 times, “he” 17 times. I noted this analysis as a possible indication of a pattern of men being addressed more often in a more respectable manner as opposed to women being addressed in the third person. It is important to go deeper into what is being said in the training data and in the case of gender bias, what are the differences between the way we refer to men and women, and how this will impact the model. Research has already showed us that ASR systems react different to the same questions depending on if a man or a woman asks them in Fessler (2017) and this analysis could help show where these

problems stem from.

Mentioned in work done by Liu et al. (2023) that analysed datasets about the importance of adopting gender-inclusive language in the training data to avoid discrimination and promote equality, I checked the Irish dataset for any gender-exclusive language. Most job titles seem to be inclusive (“author”, “photojournalist”, etc.). However there were some terms that when translated into English, using argostranslate, assumed a gender. An example of this is the gender-neutral Irish word “Cathaoirligh” got translated to “chairman” instead of “chairperson”. This indicates a problem with the translation as opposed to the dataset.

# Chapter 5

## Evaluation

After training, the model is evaluated on the validation split from the dataset. Then it is tested on the unseen data from the test split of the dataset to showcase the model's prediction capability and to calculate the metrics that will be used to compare the models with one another. This is done by executing the following steps:

1. The trained model is loaded in with a tokeniser and processor that match those used during training.
2. The dataset is filtered to contain only the test split, and the same pre-processing steps followed when preparing the data for training steps 1–4 are repeated. This involves removing unnecessary columns and special characters, and converting the audio files to the appropriate format, ensuring they are sampled at 16kHz.
3. Input values are extracted from audio samples using the processor. It then converts text labels into input IDs compatible with the model.
4. Each sample in the test dataset is iterated over and these inputs are processed:
  - (a) Inference is performed to obtain logits, which are the raw output probabilities.
  - (b) These probabilities are decoded by the processor to get the model predictions.
  - (c) The reference/ground truth text is retrieved from the test dataset.

- (d) Various evaluation metrics described in Section 5.1, such as Word Error Rate (WER), Character Error Rate (CER), Sentence Error Rate (SER) and BLEU score are calculated using the predictions and references.

### Example Prediction

Here is an example of a prediction the model that was first fine-tuned on English and then on Irish made:

**Prediction:** tá peann a gamsa agus tá peann agaita agus is linn féin éad

**Reference:** tá peann agamsa agus tá peann agatsa agus is linn féin iad

## 5.1 Metrics

All the below metrics measure an ASR system's accuracy at making predictions.

### Word Error Rate (WER)

The WER is the number of errors in the words divided by the total number of words. This is calculated using the following formula (Tomás et al., 2003):

$$WER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Total Number of Words}}$$

The “errors” in the words are defined as the number of *substitutions*, *insertions*, and *deletions* of words:

- Substitutions are essentially incorrect words, this can be through misspelling or inserting the wrong word instead of the actual word.
- Insertions are when extra words are added despite not having been said.
- Deletions are when words that were said are not detected.

A lower WER means better performance.

### Character Error Rate (CER)

Similar to WER, the CER is the number of errors in the characters divided by the total number of characters from the reference text. This is calculated using the same formula as above Equation 5.1. It is the number of substitutions, insertions, and deletions of characters divided by the total number of characters. A lower CER means better performance.

### Sentence Error Rate (SER)

SER is the percentage of sentences, whose prediction does not match the reference sentences (Tomás et al., 2003). A lower SER means better performance.

### BLEU Score

While mainly used to compare predictions to multiple reference translations, the BLEU score can still be used to measure the similarity between the prediction and the reference. It checks how many word sequences in a sentence match the word sequences of the reference (Tomás et al., 2003). A higher BLEU score means better performance.

## 5.2 Models Performance Results

The evaluation described in Section 5 was done on all the models that were fine-tuned on different languages and then on Irish. The performance results of fine-tuning these different models on Irish and using them to transcribe the Irish audio fed to that model can be seen in Table 5.1. **The Wav2Vec XLSR-53 that was fine-tuned on only Irish produces a WER of 0.662, a CER of 0.307, a SER of 0.973 and a BLEU score of 0.136.** These are the baseline model metrics and will be used to compare the other models against to see if the extra step of fine-tuning on another language dataset is worth doing. From the results below Table 5.1, we can see there is an increase and decrease in performance when transcribing the Irish audio when the model is first fine-tuned on another language dataset before being fine-tuned on Irish. For example we see lower WER, CER and SER and higher BLEU scores (which indicate a better performance)

when we first fine-tune on Arabic, Dutch, French, Persian, Portuguese and English, with English being the model with best performance boost. The model that first is fine-tuned on English has a 9.5% improvement in the WER, 7.9% improvement in the CER, 0.3% improvement in the SER and a 6.5% improvement in the BLEU score compared to the baseline model. We see worse performance when we fine-tune on German first. Performance decreases by 28.7% for the WER, 32.2% for the CER, 2.7% for the SER and 12.8% for the BLEU score compared to the baseline model. All the other models give a certain performance boost as is seen in Figure 5.1.

Table 5.1: Comparison of Model Performance

Model	WER	CER	SER	BLEU	Num of training samples
Irish	0.662	0.307	0.973	0.136	536
Arabic	0.630	0.253	0.965	0.151	14227
Dutch	0.614	0.250	0.973	0.166	9460
French	0.576	0.235	0.959	0.198	298982
German	0.949	0.629	1.000	0.008	246525
Persian	0.614	0.247	0.965	0.168	7593
Portuguese	0.590	0.241	0.954	0.180	6514
English	0.567	0.228	0.970	0.201	564337

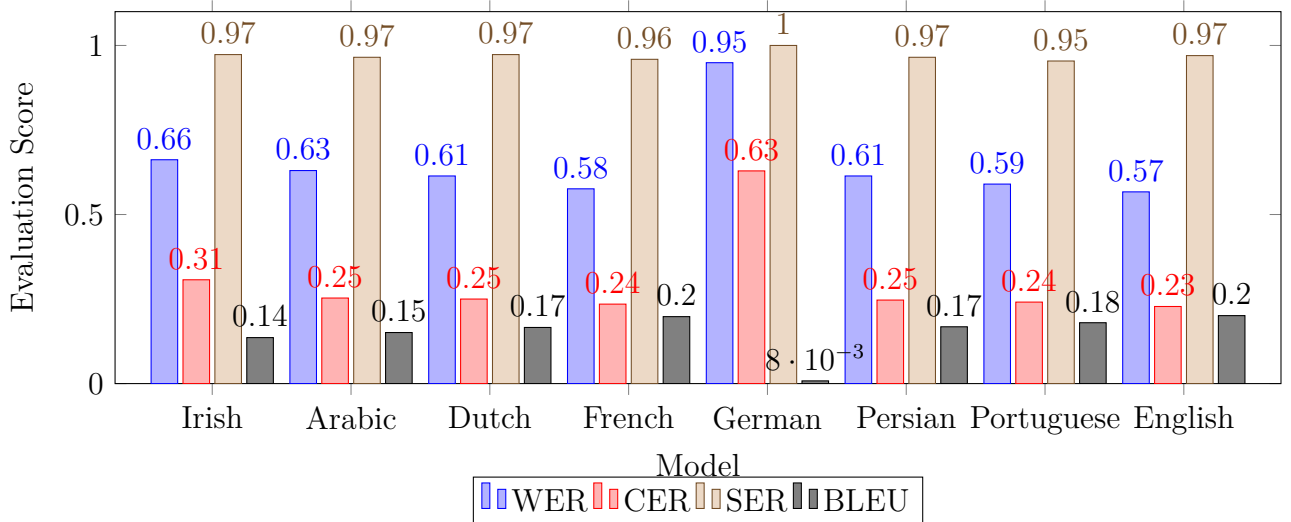





Figure 5.1: Comparison of Model Performance

There is a very strong positive correlation between WER and CER, WER and SER,

CER and SER and a very strong negative correlation between WER and BLEU, CER and BLEU and SER and BLEU shown in Table 5.2

Table 5.2: Correlation Matrix of Performance Metrics

	WER	CER	SER	BLEU	
WER	1	0.99	0.91	-0.99	 Very high correlation
CER	0.99	1	0.90	-0.97	 Strong correlation
SER	0.91	0.90	1	-0.89	 Negative correlation
BLEU	-0.99	-0.97	-0.89	1	

### Correlation between Training Samples and Performance Metric

Analysis was done on the correlation between the amount of audio clips for each language dataset used during fine-tuning, and the performance of the models fine-tuned on those datasets. What is very interesting from looking at Figure 5.2 which analyses this relationship, is that the performance of the model is not directly linked to the amount of data used for fine-tuning. While it is the English model that has the largest dataset and achieves the lowest WER rate (indicating the best performance), the French model uses a dataset almost half the size of the English one and achieves a very comparable performance. If we look at the model first fine-tuned on Portuguese, it uses a dataset 2.18% the size of the French dataset and 1.15% the size of the English dataset and there is not a huge degrade in metrics. With 98.85% less data for the Portuguese fine-tuned model we only get a 2.3% degradation in the WER. The German dataset is the third biggest dataset used for fine-tuning (after English and French) and it achieves the worst performance out of all the models for predicting Irish as indicated by the performance metrics in Table 5.1. This proves that performance is not proportionate to the size of the dataset used for fine-tuning. Although slightly better performance might be obtained sometimes by using a larger dataset, it may not be worth it for the extra time and resources that would be used. This contradicts what was concluded in Schneider et al. (2019) about using more data for pre-training having a positive effect on the performance. When fine-tuning on more data we don't see the same outcome.

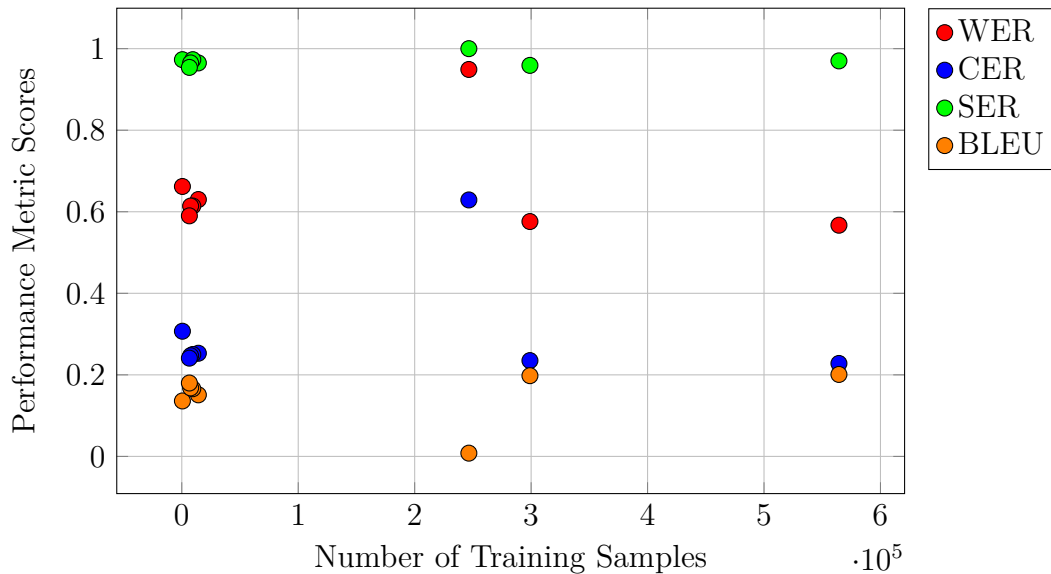


Figure 5.2: Relationship between Training Samples and Performance Metrics

## 5.3 Similarity of Datasets

To understand why certain languages were resulting in better performance than others, I followed the process recommended in Holmes et al. (2023) where the distribution of overlapping glosses are compared in order to gain an understanding of the similarities between datasets. The hope here was that it would give an indication as to which dataset is best to use to give the model a performance boost. I compared the distribution of part-of-speech (POS) tags and the distribution of word frequencies for each of the datasets used for fine-tuning, against the Irish dataset. I'll show the visual comparison of one of the language datasets in an example below, but the results of the comparisons of all the datasets will be listed at the end of the section in Table 5.3

### 5.3.1 Cosine Similarity

Cosine similarity is a measure of how similar two vectors are in a multi-dimensional space. Given two vectors in a multi-dimensional space, the cosine similarity between the two vectors is the cosine of the angle between them. The closer they are to one another in this space, the smaller the angle between the two vectors, the more similar they are



to one another. To calculate this for the POS tag and word distribution vectors for two datasets I execute the following steps:

1. I normalised the distributions by dividing each count by the total count of all items in the distribution. This ensures that the resulting values represent probabilities.
2. The magnitudes of each distribution vector are computed to make sure the cosine similarity is not thrown off by the length of the vectors. This involves summing the squares of the probabilities in the normalized distribution and taking the square root of the result.
3. I get all the unique keys (in this case the words/POS tags)
4. For each key in the set of keys, the dot product is calculated by taking the product of the corresponding probabilities from the two vectors in the normalized distributions (or 0 if the key is not present in one of the distributions). This dot product is a measure of how similar the vectors are. By summing the products of corresponding probabilities of tags between the two distributions, we measure how much the distributions overlap or align with each other in terms of the probabilities of the different tags. If two distributions have similar probabilities for the same tags, their dot product will be relatively high, indicating a high degree of similarity.
5. Cosine similarity is calculated as the dot product of the normalized distributions divided by the product of their magnitudes.

### **5.3.2 Distribution of Part-Of-Speech (POS) Tags Comparison**

Once the POS tags were gathered for each dataset, they were compared against the Irish dataset using cosine similarity described in Section 5.3.1. The result of this can be seen in Table 5.3. Here is a visualisation of the Portuguese dataset's POS tags vector plotted against the Irish dataset's POS tags vector:

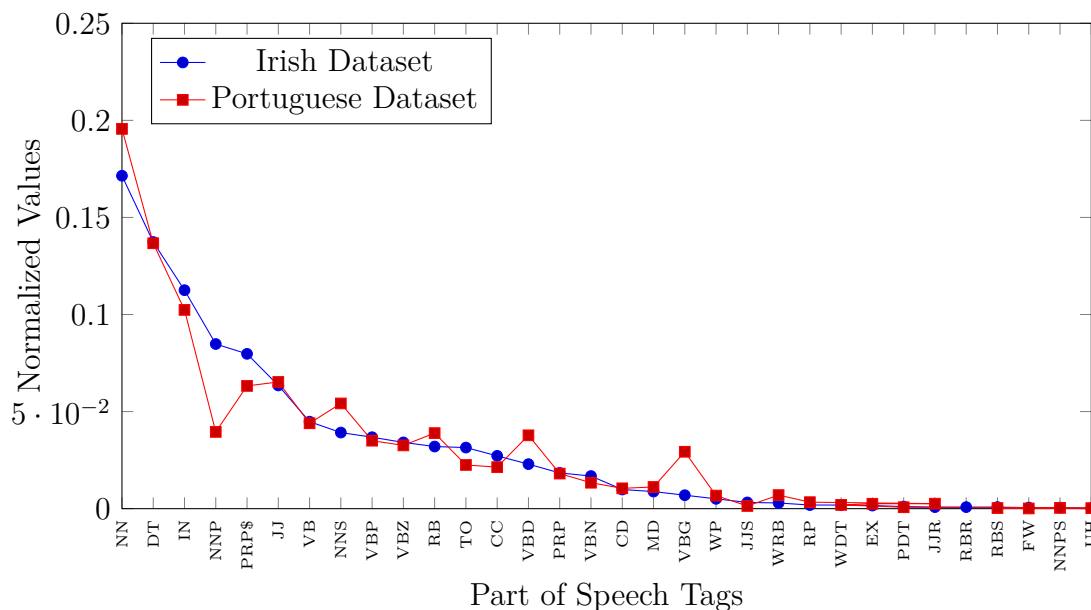


Figure 5.3: Normalized POS tagging Vector Comparison

We see in Figure 5.3 that the POS tags distributions for the Portuguese dataset do match the same vector distribution from the Irish dataset well. In general they follow a similar pattern but of course there are a few areas where they don't align. This is to be expected from two different datasets. When the cosine similarity is computed for this distribution we get a score of 0.977.

From Table 5.3 we see that the order of the datasets with closest POS tags similarity to the Irish are:

- Persian
- Portuguese
- German
- French
- English
- Arabic

- Dutch

### 5.3.3 Distribution of Word Frequency Comparison

The same process of calculating the cosine similarity for the word frequency distribution of each dataset described in Section 5.3.1 is used here. The result for all datasets can be seen in Table 5.3. Here is a visualisation of the Portuguese dataset's word frequency distribution vector plotted against the Irish dataset's word frequency distribution vector:

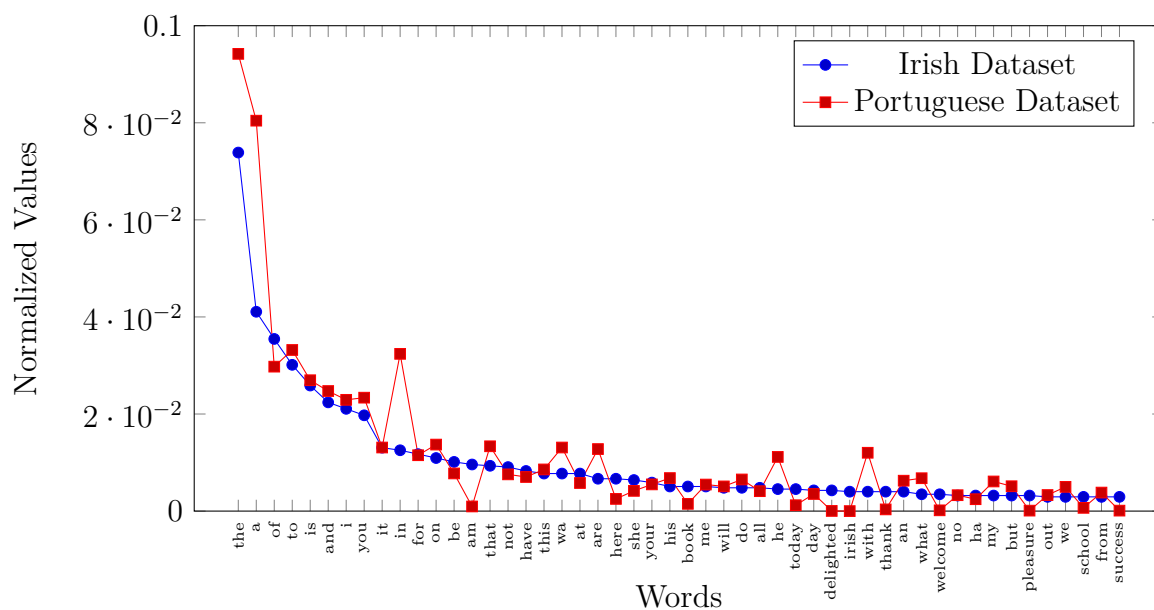


Figure 5.4: Normalized Word Distribution Vector Comparison (Top 50 words from the Irish dataset shown)

We see in Figure 5.4 that the word distribution vector from the Portuguese dataset does match same distribution from the Irish dataset reasonably well with a cosine similarity of 0.938.

From Table 5.3 we see that the order of the datasets with closest word frequency distribution similarity to the Irish are:

- Portuguese
- Persian
- English

- German
- French
- Dutch
- Arabic

### 5.3.4 Results of the Similarity of Datasets

From both of the lists above in Sections 5.3.2 and 5.3.3 we can see that this method of comparing the dataset similarities for checking which dataset is best to pre-train on does give a certain indication of which datasets will increase performance. For example the Portuguese and Persian datasets are at the top of both lists for the datasets that are closest in similarity for the word frequency distribution and the POS tagging distribution. Both of these datasets boost the performance of the model compared to the baseline and when you take into account the size of these datasets and the correlation between the number of training samples and the performance metrics in Figure 5.2, they offer the best improvements for the smallest amount of data. However this method on its own does not indicate that the German model will perform very badly and decrease the model performance. The method actually shows that the German dataset has a closer POS tag distribution similarity to Irish than English. Therefore it seems like this method on its own will not be sufficient in showcasing which dataset is best to use to improve performance.

Table 5.3: Cosine Similarity Scores

<b>Model</b>	<b>POS Tagging</b>	<b>Word Frequency</b>
Portuguese	0.9766842721296622	0.9382614286176056
Persian	0.9780358819307926	0.9214709295191432
Dutch	0.9419371717429745	0.8766793079149173
Arabic	0.9460864590582808	0.8599544997355301
English	0.9518815711270107	0.9138329463080176
French	0.9728688424256849	0.9105316896451743
German	0.9746973806519926	0.9117124326160304

## 5.4 Bias Analysis of Results

To investigate the bias that exists in these models, I decided to look at the how they perform for all groups of people in the dataset. To do this I followed the same evaluations Steps 1–4 mentioned above, but during Step 2, the dataset is filtered to only contain data from a certain group (e.g, females, males, teens, etc.). I used the WER from running this evaluation on each group to compare the performance of the groups across all the models.

### Gender

The model performs worse for women compared to men in the Portuguese 5.5, German 5.6, Persian 5.7, Dutch 5.10 and Irish models 5.12 for the WER metric. This is most likely due to:

1. The bias found in the Irish dataset against women seen in Figure 4.1.
2. The Portuguese dataset having only 4% female samples, as shown in Figure 6.1.
3. The German dataset having only 8.7% female samples Figure 6.9.
4. The Persian dataset having only 16.5% female samples Figure 6.2.
5. The Dutch dataset having no recorded female samples in the dataset Figure 6.3.

All the datasets above have a gender bias within them and this can be seen in the predictions of the model.

In the English 5.9, French 5.8 and Arabic 5.11 models, female audio from the test data performs better than male audio. Interestingly enough women perform better in the two models that were first fine-tuned on the datasets that give the best performance increase for Irish. They are English and French, Table 5.1. They performed better for females even though:

1. The French dataset has only 11.4% female samples Figure 6.10
2. The English dataset has only 2.7% female samples Figure 6.13

This indicates that not only does finding the optimal pre-training dataset increase the performance of the model but it could also help to alleviate the effects of certain biases. As for why the Arabic model performs better for women than men for the WER metric, there are studies done on Arabic ASR systems that show that these systems tend to exhibit bias against men with an average WER of 1.33% for women whereas male speakers obtain an average WER of 2.29% (Sawalha and Abu Shariah, 2013). Also the Arabic dataset is the only dataset I analysed where there are more females recorded in the data than males Figure 6.12.

## Age

Teens perform consistently better in the Portuguese, German, French, Dutch and Arabic models. Teens are the age group with the second lowest amount of labelled data associated with it for Irish dataset, Figure 4.2. Interestingly in the other datasets teenagers are:

1. The age group with the smallest representation in the Portuguese dataset as seen in Figure 6.4
2. The age group with the smallest representation in the German dataset from the age groups that we tested on Figure 6.11.
3. The age group with the smallest representation in the French dataset from the age groups that we tested on Figure 6.5.
4. Not recorded at all in the Dutch dataset Figure 6.6.
5. And again age group with the smallest representation in the Arabic dataset Figure 6.7.

Teenagers don't perform best in the Persian, English and Irish model. In these models the 40s age group performed best across all of them even though:

1. There was only 1 speaker in their 40s labelled in the Irish dataset Figure 4.2.
2. 40s was the third highest group on speakers in the English dataset Figure 6.14

3. 40s were the second lowest recorded age group in the Persian dataset Figure 6.8

This proves the amount of data for an age group does not directly correlate to the performance. It also ties in well with the idea mentioned in Lonergan et al. (2023b) that balanced corpora can result in unequal performance and that we need to investigate further as to why bias exists.

On further analysis of why I was seeing better performance of the voices of, I looked into the reasons behind the accents of young people being easier to be classify. I were trying to find out if there was any kind of neutralisation of dialects happening for younger generations. This theme is seen across two papers that look at newer, younger accents for English in Ireland and Australia. Korhonen (2017) looks at the American influence on Australian English and people's perception of it. They found that people did notice that dialects in Australia were being Americanised and that "younger speakers are more likely to use the more American style variants" (Korhonen, 2017). A similar pattern was seen by Moore (2011) when inspecting the "D4" accent in Irish-English. They found that younger people today were more inclined to take on this accent and its "quasi-American inflection" with the modernisation of Dublin in recent years. This has given the young Irish speakers of English a want for "urban sophistication", which entails moving further away from their roots and the dialects of older generations and more towards and this Americanised way of speaking. A quote from a journalist in Moore (2011) sums this up nicely "Across the country, people are abandoning their regional lilt and embracing that flat quasi-American inflection colloquially known as the Dublin 4 accent, according to language experts. As a result, the traditional Irish brogue - and its many variations - may be in mortal danger." (Moore, 2011). This leads me to think the same trend could be seen across speakers of the Irish language. The dialects could be becoming more "neutral" amongst younger people. This would explain why the language models performs better on teens as this Americanisation of dialects may be seen across other languages in the datasets used during pre-training, thereby boosting performance.

## Dialect

Gaeilge Uladh performs best across the Portuguese 5.5, German 5.6, Persian 5.7, English, Dutch 5.10, Arabic 5.11 and Irish model 5.12. From the analysis on the Irish dataset we saw that the Gaeilge Uladh dialect has the second most amount of audio associated with it, but Gaeilge na Mumhan only falls slightly behind by 3 audio clips Figure 4.3. Therefore this doesn't really give a good reason as to why there is a bias towards the Uladh dialect in performance. Perhaps the better performance is due to the fact that the Uladh dialect is more distinct from the other two accents which models seem to have difficulty disambiguating between (Lonergan et al., 2023b) This behaviour of the Uladh dialect performing better was also seen in Lonergan et al. (2023b) after a pre-trained ECAPA-TDNN was fine-tuned on Irish.

The French model 5.8 is the only model where the Uladh dialect does not perform the best, however it only performs 1.35% behind the Gaeilge na Mumhan dialect. Gaeilge Chonnacht performs the worst across all models even though it is the dialect with the most labelled audio data in the Irish dataset Figure 4.3. This again confirms that more data per group of speakers, does not result in better performance. There is another explanation here for why the English and French datasets align well with the Irish dataset and therefore provide better performance. There is evidence of both of these languages having an influence on Irish and alongside Latin, English and French are the main languages in which Irish loaned words (Karkishchenko, 2010). It was French that was mainly spoken around towns in Ireland during the twelfth century after the Anglo-Norman invasion, and then it gradually moved towards English by the thirteenth century.

## Best Models by Group

Below in Table 5.4 we have the different groups that I tested the models on, and which model that was first fine-tuned on a particular dataset performed best for that group. From Table 5.4 it is clear that the model first fine-tuned on English performs the best across most groups (females, males, teens, 20s, 40s, Gaeilge Uladh and Gaeilge Chon-



nacht). However there are certain groups where other models perform better, such as people in their 50s where the Persian model performs best and people with a Gaeilge na Mumhan dialect where the French model performs best. These kind of peculiarities were interesting to note, since for example people in their 50s were the lowest labelled age group in the Persian model Figure 6.8, and there were no people with a Gaeilge na Mumhan dialect in the French dataset. This suggests that analysing the labels on the dataset alone is not enough to get to the bottom of why models behave as they do and it is not the best way to eliminate bias. We might have to look at the pitches of voices and the sounds made by different dialects to understand what gives certain groups a better performance.

Table 5.4: Best Performing Models for Each Group based on WER

Group	Best Model (WER)
Females	English Model
Males	English Model
Teens	English and French Model
20s	English Model
30s	Persian and Dutch Model
40s	English Model
50s	Persian Model
Gaeilge Uladh	English Model
Gaeilge Chonnacht	English Model
Gaeilge na Mumhan	French Model

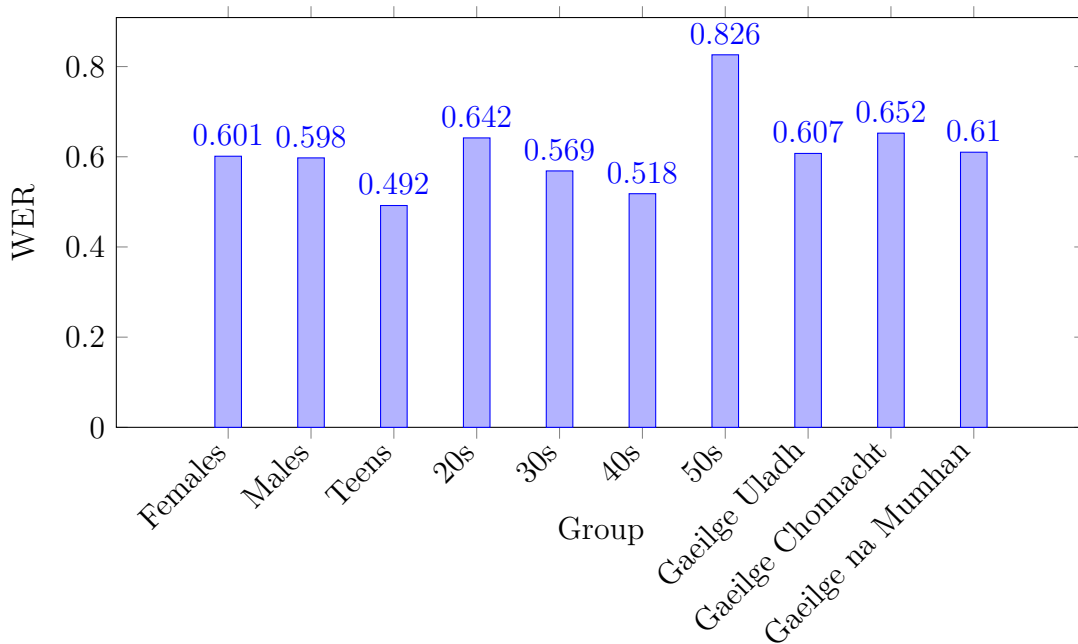


Figure 5.5: WER by Group Portuguese Dataset

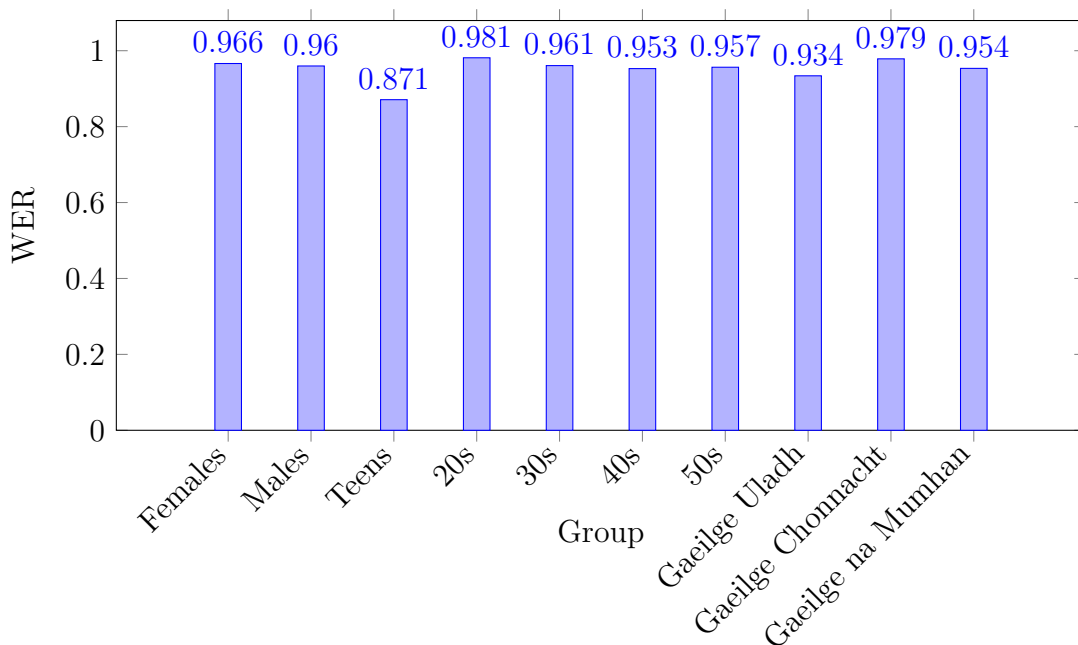


Figure 5.6: WER by Group German Model

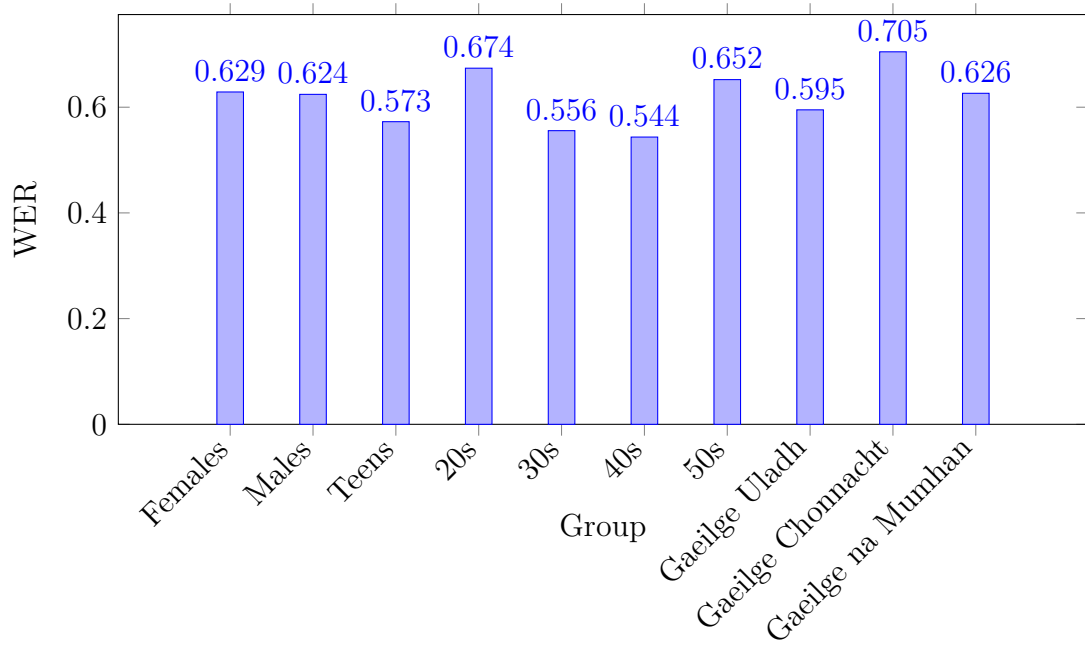


Figure 5.7: WER by Group Persian Model

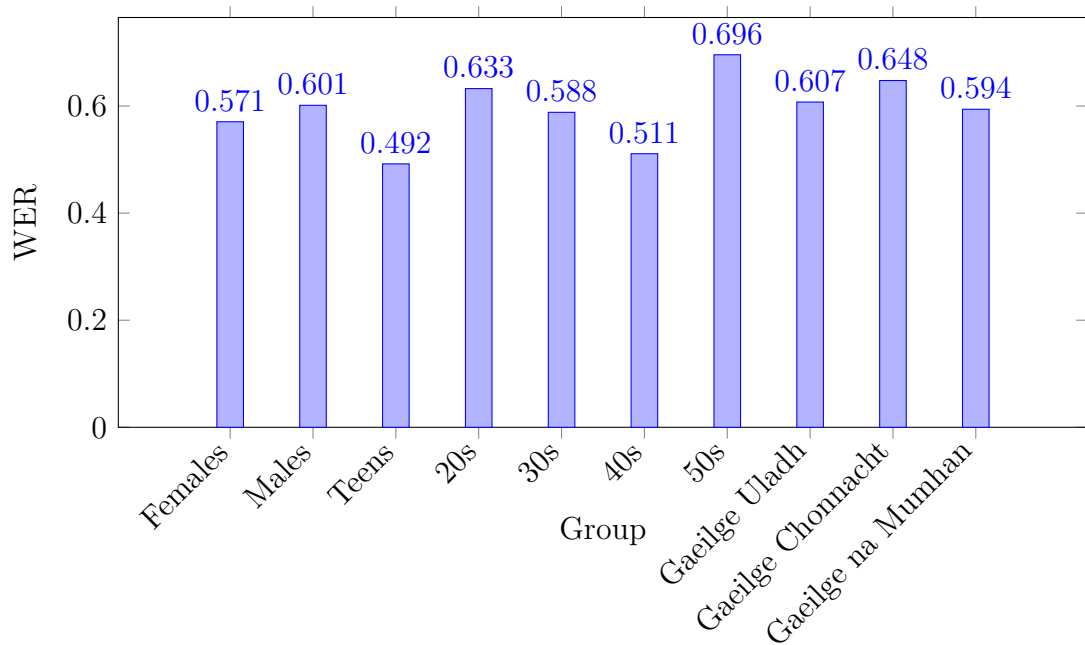


Figure 5.8: WER by Group French Model

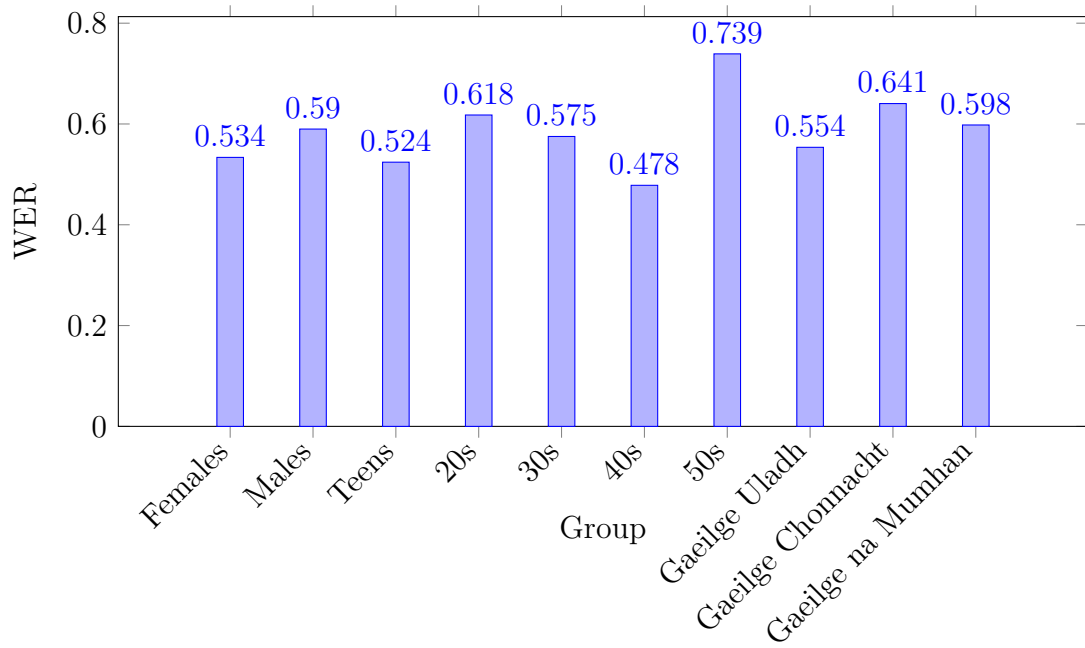


Figure 5.9: WER by Group English Model

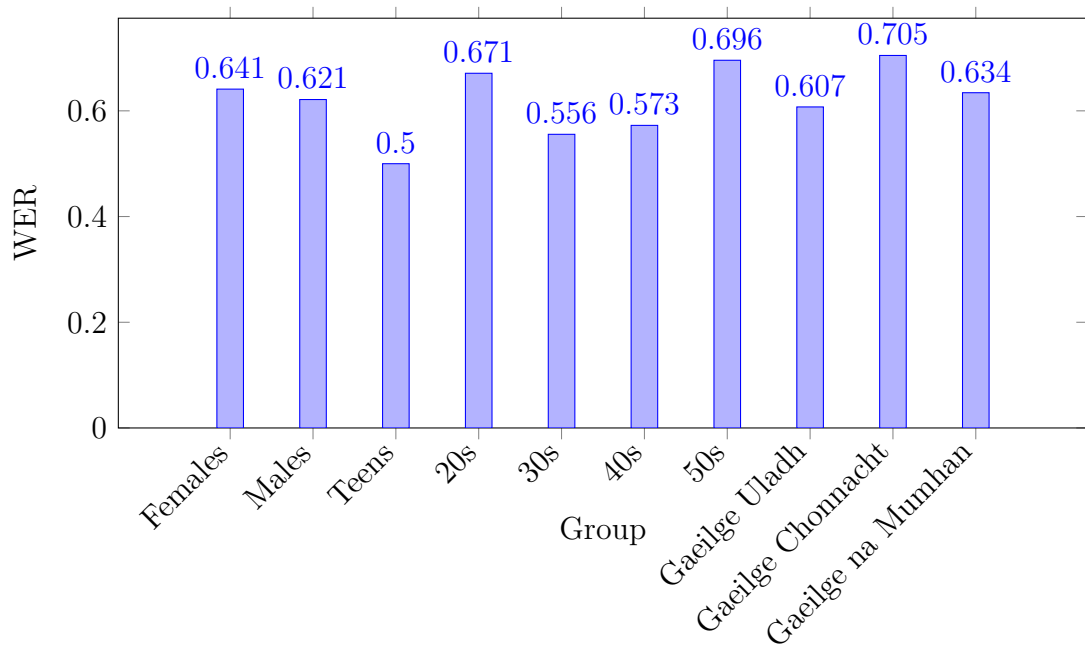


Figure 5.10: WER by Group Dutch Model

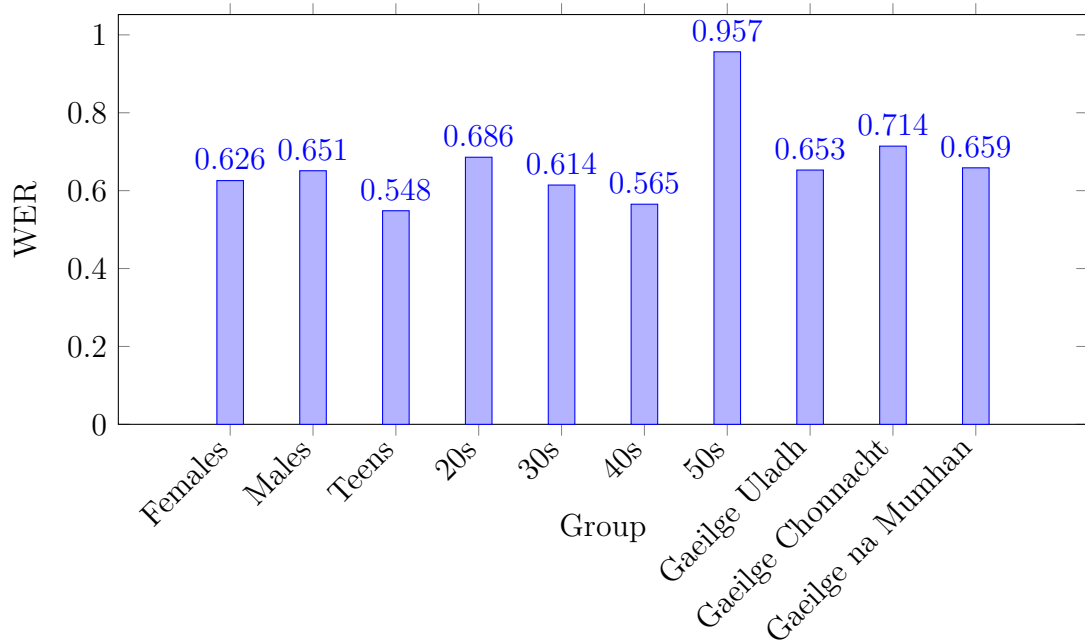


Figure 5.11: WER by Group Arabic Model

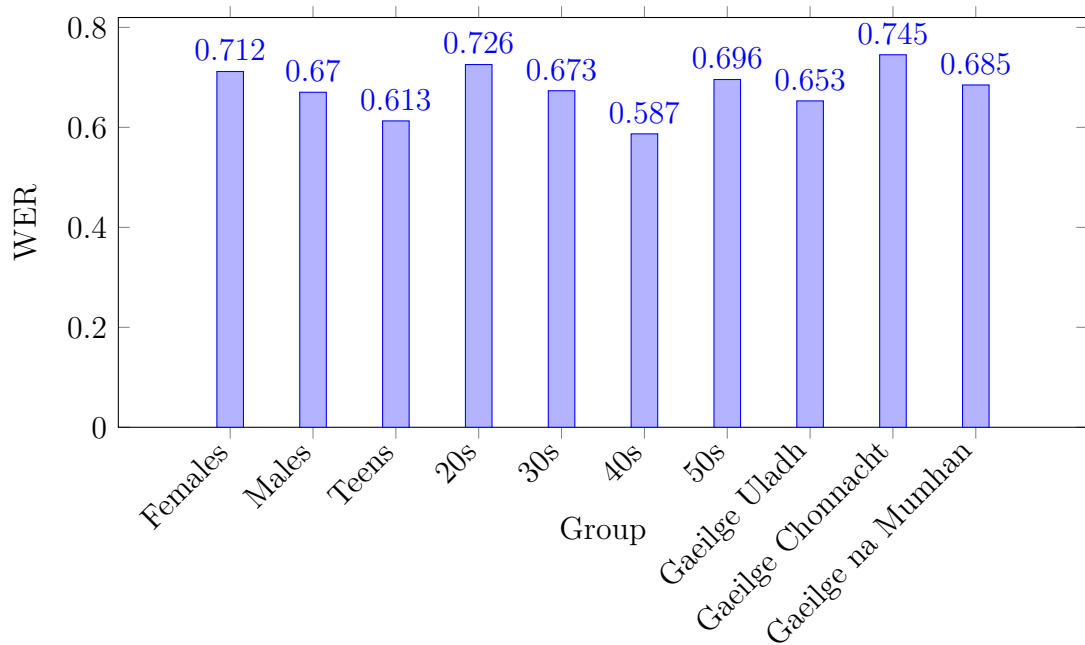


Figure 5.12: WER by Group Irish Model

# Chapter 6

## Conclusions & Future Work

The intention of this paper was to find out about the challenges the Irish language faces, what biases exist in an Irish open source dataset, does gloss analysis help identify which language will give an Irish ASR systems the best performance boost and how does transfer learning boost the performance of these systems. I have found that Irish faces many problems as a low-resource language from problems with ASR systems understanding and identifying dialects due to it not having a “spoken standard” to a lack of research and funding resulting in learners and speakers of the Irish language not having the same quality of technology available to them as other languages do.

The bias within the Irish Common Voice dataset was also investigated and it was found that it does contain an over-representation of certain voices and an under-representation of others, particularly females and people over the age of 60. The bias that exists in all ASR systems was researched and the root of the problem was found to be the datasets used to train these ASR systems, therefore I was expecting that the model I created using the Irish Common Voice dataset which I deemed to be biased, would result in a biased model. This was proven by the unequal performance of this model on certain groups of people 5.12. The performance of the model could at times be linked back to the datasets used for fine-tuning particularly when it can to bias in the gender performance, however not all the performance bias could be linked back to the language datasets that were used to fine-tune the models. Even when there was a good representation for a certain group in the dataset,

this would not translate to a good performance for that group in the model. Therefore we can not conclusively say that balanced corpora are the solution to eliminating bias as bias is not solely related to representation. We need to further investigate the contents on the data in the dataset and perhaps linguistic characteristics such as the pitch of voices. A deeper understanding of the data is needed to understand the models predictions, proven by model performance being linked to the neutralisation of younger speakers accents and the influence of French and English on the Irish language. We did find that transfer learning, in my case the transferring of learning from fine-tuning on a certain language before fine-tuning on Irish, did improve the speech recognition capabilities of the model compared to the baseline model by 9.5% for the WER metric when the English dataset was used for the first round of fine-tuning. However, on further investigation into the different datasets it was found that there is a weaker correlation between the size of the dataset used for fine-tuning than expected and the performance of a model, which is an interesting finding especially when trying to reduce the environmental impact of these models. When using the Portuguese dataset instead of the English one with 98.85% less data for fine-tuning, we only get a 2.3% degradation in the WER metric, providing an impressive 7.2% performance boost in the model compared to the baseline. This proves that fine-tuning is definitely worth doing but in some cases when the wrong dataset is used it can lead to a decrease in performance.

When using gloss analysis to determine which dataset was best to initially fine-tune the model on for the best performance boost, we found that it did give some indication as to which datasets might work best, but it can't be relied on to determine a dataset that will decrease the performance of the model and will not provide a ranking of the datasets that could be used. It can't be used to decipher a good dataset for pre-training from a bad one.

## **6.1 Future Work**

While doing this work it was clear that a lot of open source datasets lack the required metadata to thoroughly understand the characteristics of a dataset and to clearly see the distributions of different groups of people. Without this information it is very difficult to know where exactly the bias in training data comes from and can have damaging affects, resulting in discrimination towards certain groups of people. Work needs to be done to label this data that many people are accessing and building ASR systems with.

An interesting find from this paper was that there was a lower correlation than expected between the size of the dataset and the performance of a model. When fine-tuning with the German dataset, I found that it negatively impacted the performance of the model. Work has been done by Solaiman and Dennison (2021) on curating datasets to achieve a better and more unbiased performance when fine-tuning. It would be interesting to see if a dataset that when initially used for fine-tuning, resulted in a poor model performance, could be curated to achieve a better performance. In my case it would be interesting to see if the German dataset could be curated, so that using it resulted in a better performance than the English dataset. As mentioned above, work needs to be done on the subtle links in data that leads to certain groups usually having a better performance such as Gaeilge Uladh or male speakers. This could involves a more in-depth investigation into the sounds and pitches of dialects and speakers.

This research could be extended into looking at models that are not fine-tuned on the datasets and rather trained from scratch on the datasets to try an pin point exactly where the bias in the performance is coming from.



# Appendix

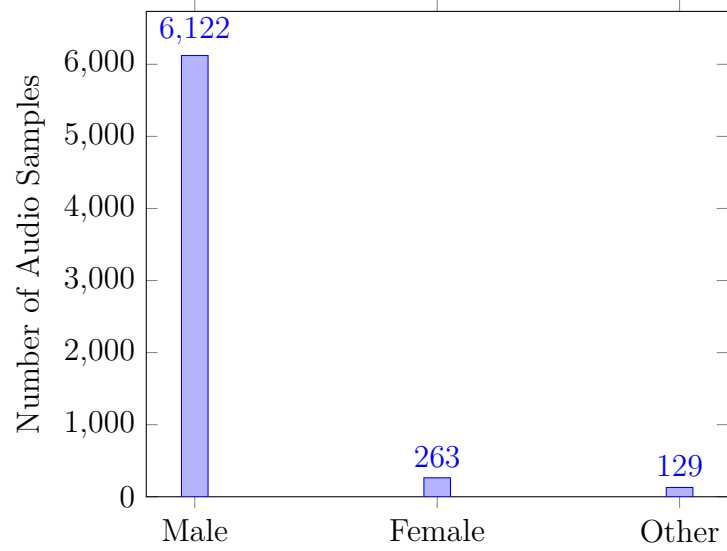


Figure 6.1: Gender Distribution Portuguese Dataset

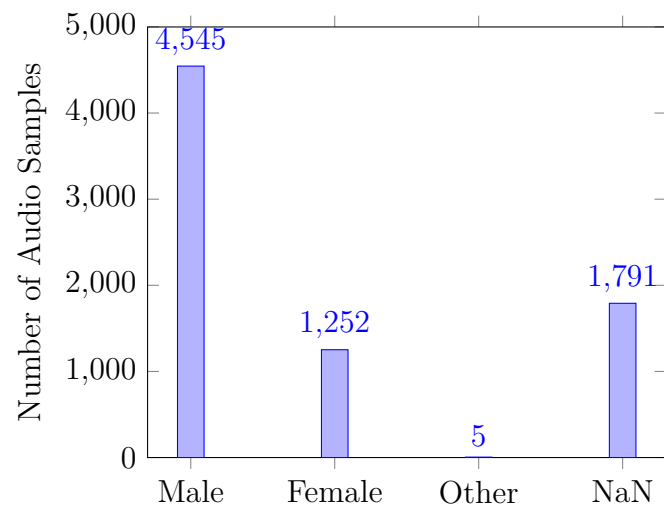


Figure 6.2: Gender Distribution Persian Dataset

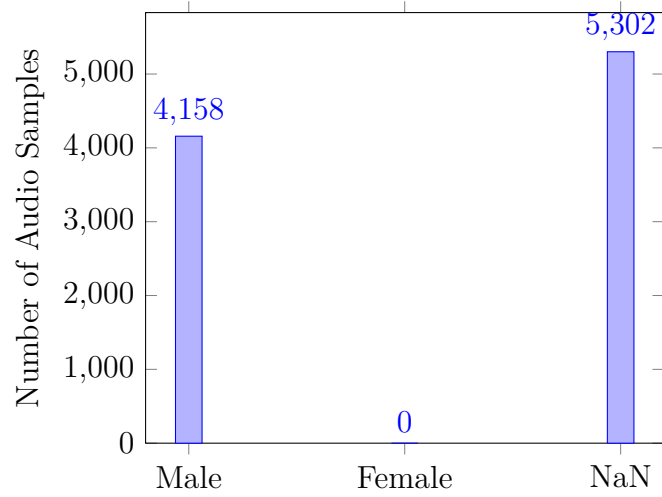


Figure 6.3: Gender Distribution Dutch Dataset

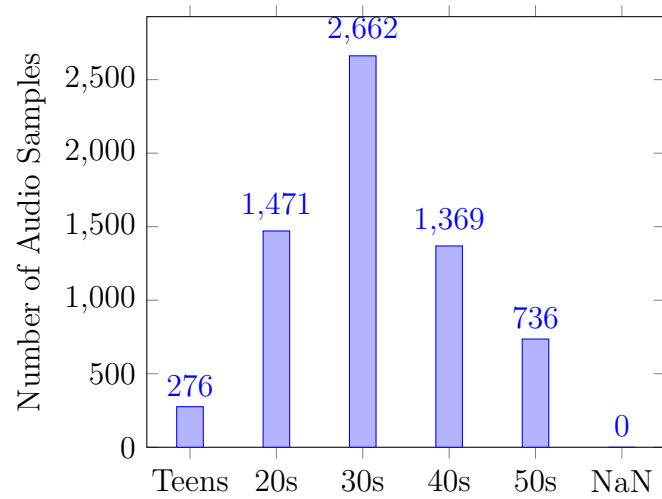


Figure 6.4: Age Distribution Portuguese Dataset

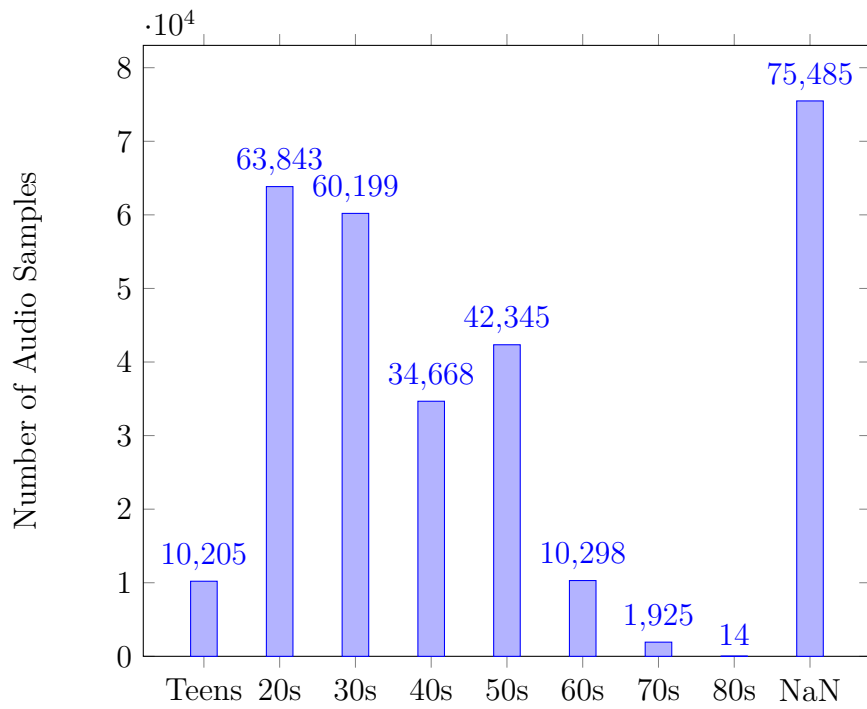


Figure 6.5: Age Distribution French Dataset

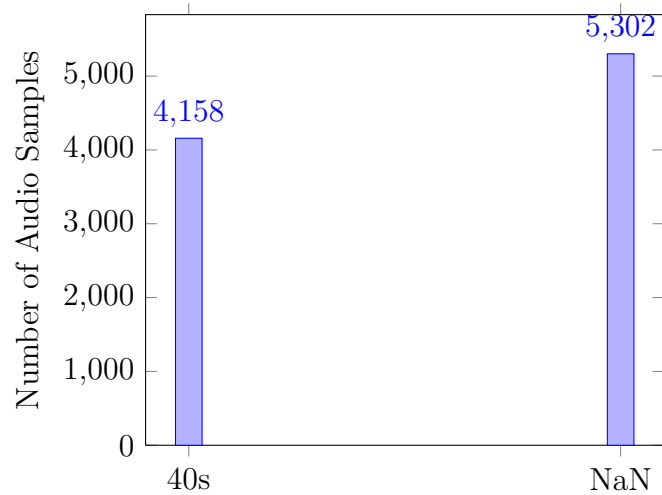


Figure 6.6: Age Distribution Dutch Dataset

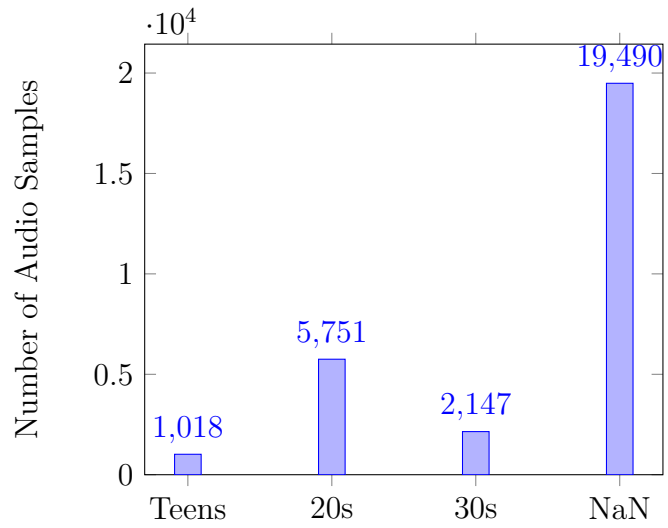


Figure 6.7: Age Distribution Arabic Dataset

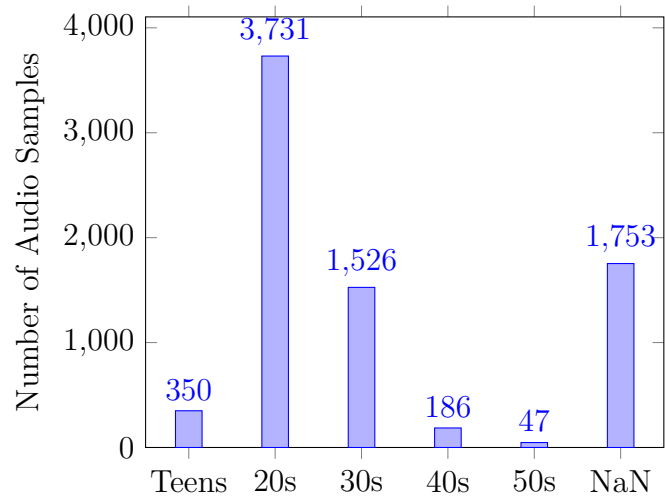


Figure 6.8: Age Distribution Persian Dataset

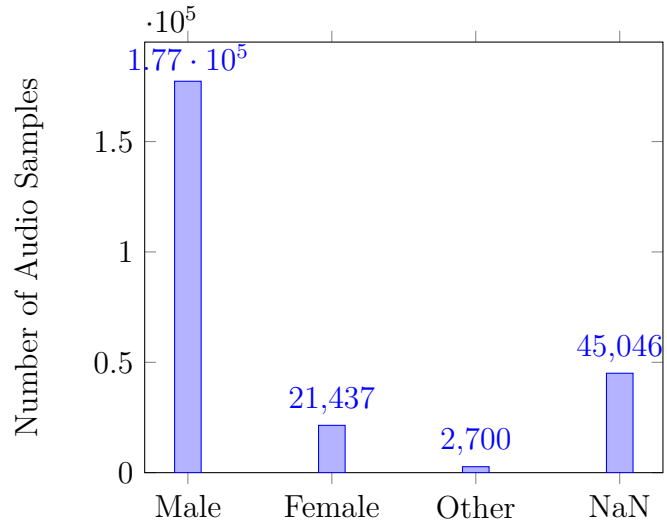


Figure 6.9: Gender Distribution German Dataset

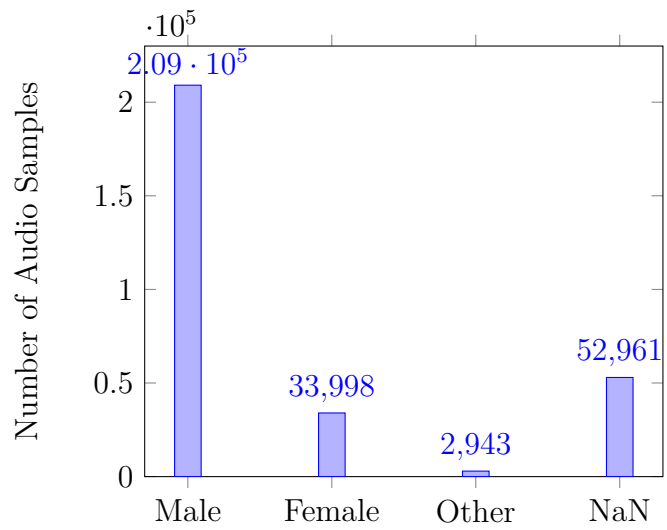


Figure 6.10: Gender Distribution French Dataset

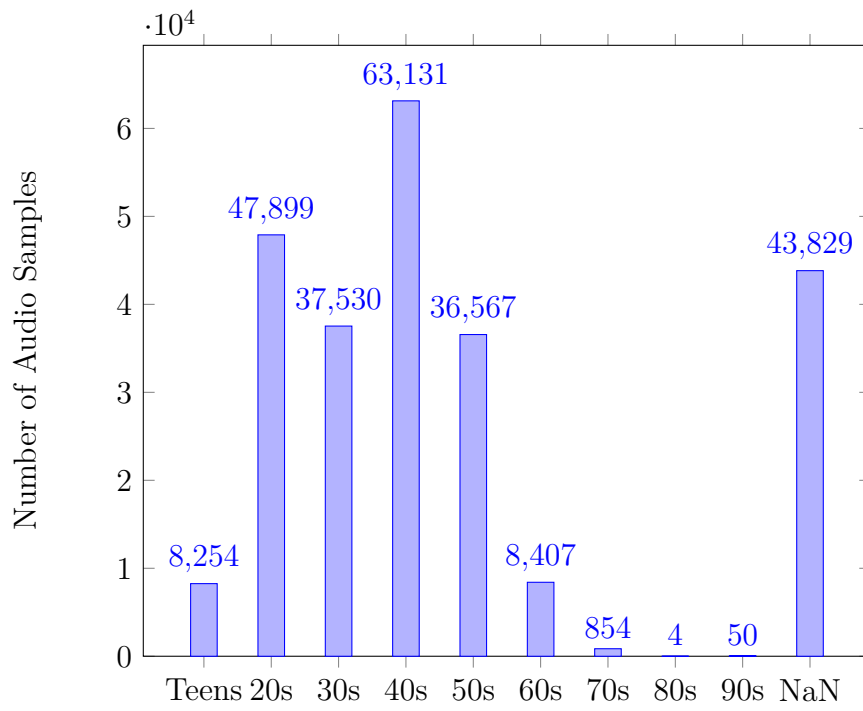


Figure 6.11: Age Distribution German Dataset

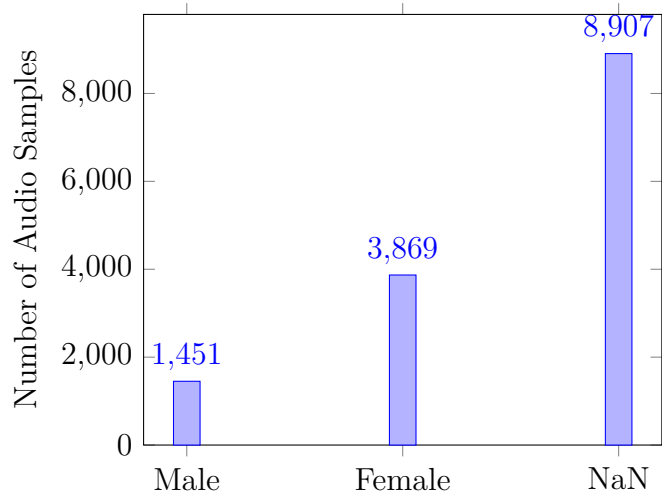


Figure 6.12: Gender Distribution Arabic Dataset

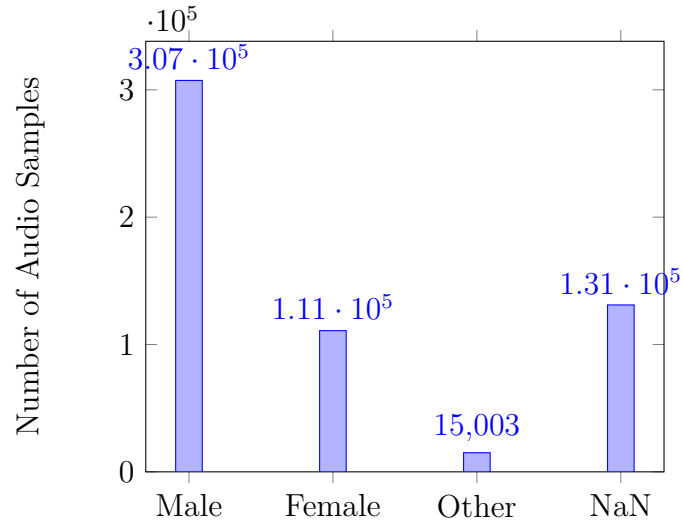


Figure 6.13: Gender Distribution English Dataset

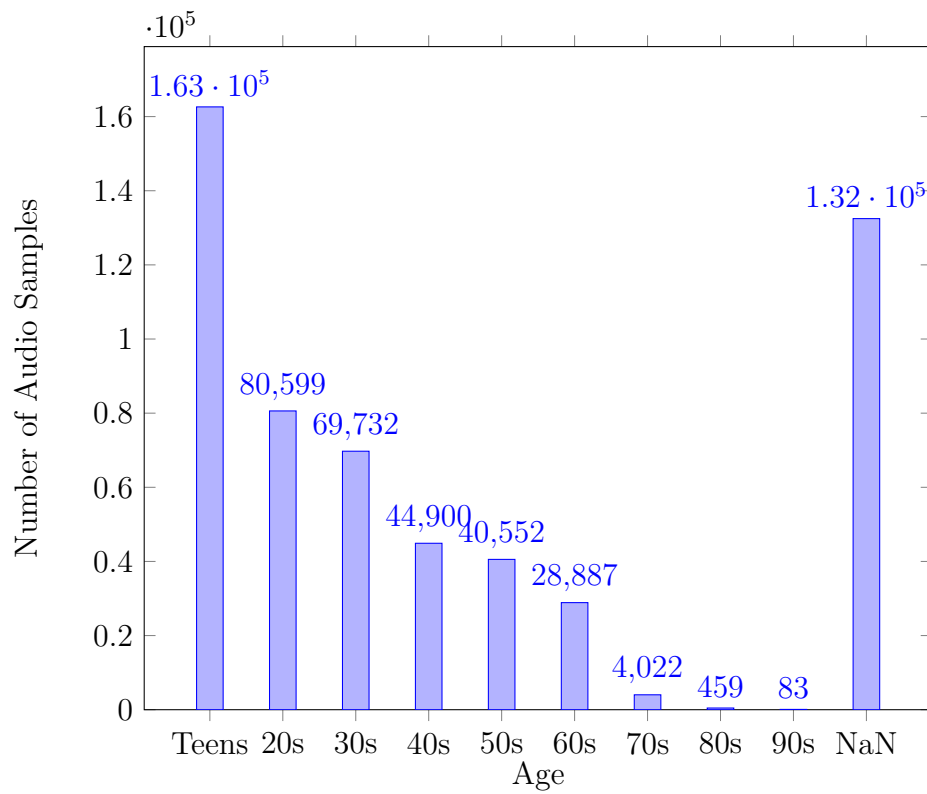


Figure 6.14: Age Distribution English



# Bibliography

- Adda, G., Stüker, S., Adda-Decker, M., Ambouroué, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., et al. (2016). Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14.
- Amazon.com, Inc. (2024). Alexa. URL: <https://developer.amazon.com/en-US/alexa>.
- Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Barnes, E., Morrin, O., Chasaide, A. N., Cummins, J., Berthelsen, H., Murphy, A., Corcráin, M. N., O’Neill, C., Gobl, C., and Chiaráin, N. N. (2022). Aac don ghaeilge: the prototype development of speech-generating assistive technology for irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 127–132.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Central Statistics Office (Ireland) (2016). Census of population 2016 – profile 10 education, skills and the irish language. Retrieved from <https://www.cso.ie/en/releasesandpublications/ep/p-cp10esil/p10esil/ilg/>.

Chiaráin, N. N., Nolan, O., Comtois, M., Gunning, N. R., Berthelsen, H., and Chasaide, A. N. (2022). Using speech and nlp resources to build an icall platform for a minority language, the story of an scéalaí, the irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 109–118.

Chiblow, S. and Meighan, P. J. (2022). Language is land, land is language: The importance of indigenous languages. *Human Geography*, 15(2):206–210.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fessler, L. (2017). We tested bots like siri and alexa to see who would stand up to sexual harassment. *Quartz Magazine*.

Gales, M. J., Knill, K. M., Ragni, A., and Rath, S. P. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Garofolo, J. S., Graff, D., Paul, D., and Pallett, D. (1993). Csr-i (wsj0) complete ldc93s6a. *Philadelphia: Linguistic Data Consortium.*, page Web Download.

Glasser, A., Kushalnagar, K., and Kushalnagar, R. (2017). Deaf, hard of hearing, and hearing perspectives on using automatic speech recognition in conversation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 427–432.

Google LLC (2024). Google home. URL: <https://home.google.com/welcome/>.

- Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.
- Grosman, J. (2021). Fine-tuned XLSR-53 large models for speech recognition. <https://huggingface.co/jonatasgrosman/>.
- Harwell, D. (2018). The accent gap. *The Washington Post*.
- Holmes, R., Rushe, E., De Coster, M., Bonnaerens, M., Satoh, S., Sugimoto, A., and Ventresque, A. (2023). From scarcity to understanding: Transfer learning for the extremely low resource irish sign language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2008–2017.
- Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.
- Karkishchenko, O. (2010). Anglo-norman borrowings in irish.
- Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2021). Racial disparities in automated speech recognition. In *Proc. of the National Academy of Sciences*.
- Korhonen, M. (2017). Perspectives on the americanisation of australian english.
- Liu, C., Jhang, S., and Lee, S. (2023). From gender-biased to gender-specific and gender-inclusive words: A corpus-based study. , 31(1):85–112.
- Lonergan, L., Qian, M., Berthelsen, H., Murphy, A., Wendler, C., Chiaráin, N. N., Gobl, C., and Chasaide, A. N. (2022). Automatic speech recognition for irish: the abair-éist system. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51.

- Lonergan, L., Qian, M., Chiaráin, N. N., Gobl, C., and Chasaide, A. N. (2023a). Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer? *arXiv preprint arXiv:2307.07295*.
- Lonergan, L., Qian, M., Chiaráin, N. N., Gobl, C., and Chasaide, A. N. (2023b). Towards spoken dialect identification of irish. *arXiv preprint arXiv:2307.07436*.
- Lynn, T. (2023). Language report irish. In *European Language Equality: A Strategic Agenda for Digital Language Equality*, pages 163–166. Springer.
- Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.
- Martin, J. L. and Tang, K. (2020). Understanding racial disparities in automatic speech recognition: The case of habitual” be”. In *Interspeech*, pages 626–630.
- Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., and Tuennerman, E. (2021). “i don’t think these devices are very culturally sensitive.”—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4:725911.
- Moore, R. (2011). ” if i actually talked like that, i’d pull a gun on myself”: accent, avoidance, and moral panic in irish english. *Anthropological Quarterly*, pages 41–64.
- Mozilla Corporation (2021). Mozilla Common Voice. <https://commonvoice.mozilla.org>. Accessed: 29th September 2023.
- Murtagh, L. (2003). Retention and attrition of irish as a second language. *Unpublished PhD Thesis, University of Groningen*.
- Ngueajio, M. K. and Washington, G. (2022). Hey asr system! why aren’t you more inclusive? automatic speech recognition systems’ bias and proposed bias mitigation techniques. a literature review. In *International Conference on Human-Computer Interaction*, pages 421–440. Springer.

- Park, K. and Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *Interspeech*.
- Quéniart, A. and Charpentier, M. (2012). Older women and their representations of old age: a qualitative analysis. *Ageing & Society*, 32(6):983–1007.
- Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.
- Sawalha, M. and Abu Shariah, M. (2013). The effects of speakers’ gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds.
- Schneider, S., Baeovski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
- Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.
- Tithe an Oireachtais (2017). *Gramadach na Gaelige, An Caighdeán Oifigiúil*. Seirbhís Thithe an Oireachtais.
- Tomás, J., Mas, J. À., and Casacuberta, F. (2003). A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 27–34.
- Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global.

Yang, L. and Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316.