PRIFYSGOL
BANGOR
UNIVERSITY

School of Arts, Culture and Language
College of Arts, Humanities and Business

Transfer Learning for Speech-to-Text: Investigating the Impact of the Base Language on the Performance of Models

XXX

Submitted in partial satisfaction of the requirements for the
Degree of Master of Science
in Language Technologies

Supervisor Dr. William J. Teahan

September 2022

Abstract

This dissertation explains the work that was undertaken to investigate the fac- tors that affect the performance of models trained using transfer learning. It outlines how the relationship between two languages were quantified and how this data was used to derive several quantifiable metrics to be used for the analysis. The dissertation then explains how several scripts and tools were developed to enable transfer learning to be used in the process of training a range of models. The models combined with the metrics extracted from the relationships enabled different factors that affect the performance of the models to be analysed. In total, 29 models were trained and 25 of these were trained for lower-resourced languages. State-of-the-art models were achieved for Breton and Romansh, while the first monolingual Galician models were trained. Effec- tive models comparable to the state-of-the-art were also trained for both Welsh and Portuguese. While the findings could not definitively show any correlations between the performance of the models and the relationship between the base and target language, other factors were uncovered and exhibited statistically sig- nificant correlations. The dissertation shows that there is a correlation between a base model's ability to perform its own learning task and the performance of the models that used this model as a base.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Good acoustic models require large amounts of data, and there is often a relationship between the amount of training data and the performance of models. This is a problem for lower-resourced languages as data is expensive and hard to come by, especially spoken data, and the data required to make language models is often quite substantial. However, with new techniques, there are op- portunities to create base models based on a different language and use these as a foundation on which further training can be done. This process is called transfer learning. The idea is that these bilingual and multilingual models can be used to create models for lower-resourced languages in situations where there is not enough data available to create a model from scratch.

This dissertation seeks to investigate transfer learning for speech-to-text in greater depth. It will attempt to evaluate the viability of this approach and how it compares to other methods. It will also investigate what factors contribute to the quality of speech-to-text models and attempt to determine what role the selection of language has on the overall quality of the models.

1.1 Background and Motivation

Lower-resourced languages like Welsh and Breton have long struggled with lim- ited available data when trying to create effective acoustic models. A digital presence and effective models are vital for lower-resourced languages and for ensuring that they do not become digitally extinct. However, many languages simply do not have the available data to produce effective models from scratch. It is therefore of vital importance to investigate in what ways the limited data can be utilised as effectively as possible. By improving the optimisation of data utilisation, the barrier of entry for technologies like speech-to-text can be lowered, enabling a range of languages to benefit from these technologies.

While the factors that contribute to effective transfer learning models have been explored in other domains, this has not been explored in-depth in the context of speech-to-text. Since acquiring speech data is expensive and time-

consuming, this is especially problematic for lower-resource languages since they have low amounts of data and requires effective data utilisation. For example, one area where existing literature falls short is in answering what effect the choice of base language has on the effectiveness of the transfer learning process. The choice of base language has been shown in other domains to play a not insignificant role in determining the effectiveness of transfer learning.

Attempting to answer these questions and explore ways to optimise the effectiveness of transfer learning is therefore an important step toward improving models for lower-resourced languages. Achieving this and ensuring that lower resources languages are able to better utilise the data that is available to them remain the underlying motivations for this dissertation.

1.2 Aim and Objectives

The overarching aim of this dissertation is to improve speech-to-text models for lower-resourced languages and to explore ways of optimising data utilisation for transfer learning. To achieve this, the dissertation aims to answer some unanswered questions in regard to what impact the relationship between the base language and target language has on the effectiveness of transfer learning. By investigating this, the dissertation aims to uncover whether it is possible to improve the utilisation and effectiveness of available data for minority and lower- resourced languages. By doing this, the dissertation aims to lower the barrier of entry for these languages enabling them to have effective speech-to-text models and to thrive in the digital world despite their lower-resourced status.

To achieve this aim, several concrete objectives will have to be achieved.

These objectives are:

1. Create a way of extracting information from the training set to quantify the relationship between two languages.
2. Train bespoke and novel models for several lower-resourced languagesus- ing transfer learning.
3. Investigate whether there is a correlation between the relationship between the languages and the performance of the models.
4. Explore whether there are any other contributing factors that affect the performance of the models.

1.3 Contributions

This dissertation has answered several unanswered questions in relation to trans- fer learning and uncovered some potential contributing factors as to what makes transfer learning effective. The findings in this dissertation enable transfer learn- ing to be more effectively utilised, and for more effective models to be developed

in the future. This has the effect of lowering the barrier of entry to speech-to- text technologies for lower-resourced languages.

The dissertation produced a wide range of speech-to-text models for a total of five languages: Welsh; Breton; Romansh; Galician; and Portuguese. In addition to these, two models were produced for both French and German. In total, the dissertation produced four base models and 25 target models. Most of the models were comparable with the available state-of-the-art models, with many of the models also outperforming the state-of-the-art significantly. Major improvements were achieved for speech-to-text models for Breton and Romansh, and what appears to be the first monolingual speech-to-text models for Galician were developed.

The dissertation also shows that by using transfer learning effectively, small teams with limited resources can effective speech-to-text models that are usable by their communities. The methodology laid out in this dissertation providesa tangible way for effective models to be developed for smaller and lower-resourced languages. This is a real impact on enabling these languages to survive and thrive in the digital world and helps prevent them from going digitally extinct.

1.4 Summary of Dissertation

The next chapter provides an overview of existing literature, technologies, and advances within the field. By exploring the state-of-the-art frameworks within speech-to-text and transfer learning, the chapter highlights where existing re- search falls short and why the topic that this dissertation covers is an area that should be investigated further.

Chapter 3 will formulate the research questions and hypotheses that this dissertation will investigate. The methodology that the dissertation will use to investigate these will also be described including definitions for any required metrics.

Chapter 4 will describe any work that was undertaken to preprocess the data before any of the models are trained. An analysis will be carried out into how well Common Voice utilises the available data, and determine whether this needs to be improved. Then we will explore how we can extract metrics from the training data in an attempt to measure how close two languages are to each other phonetically. These metrics can then later be used to analyse whether there are correlations between these metrics and the performance of the models.

In chapter 5, we will use transfer learning to create a set of novel models for Welsh and Breton. To achieve this a training environment and several scripts have to be created to aid the training process. Following this, several several base models will be created for both English and French. Using these models a total of three models for both Welsh and Breton can be created.

Using the metrics extracted in chapter 4, the performance of these models will then be analysed and we will see whether there is correlation between these metrics and the performance of the models. We will also consider otheraspects

of the transfer learning process to investigate whether there is anything else that we can learn about this process.

Due to the results uncovered in chapter 5, chapter 6 will expand the experi- ment by including a set of new languages; German, Romansh, Portuguese, and Galician. Using the additional data gather by increasing the sample size, we are able to undertake a more robust analysis than what was possible in the original analysis.

The final chapter will summarise the findings, discuss some of the impli- cations of these, and discuss topics that should be investigated further in the future.

Chapter 2

Literature Review

In this chapter, we will review the existing literature and the current state-of- the-art speech-to-text (STT) systems. In addition to this, we will look at how existing literature falls short of answering some important questions in regard to how transfer learning for speech-to-text could be further optimised. We will also review existing resources that provide a basis for speech-to-text systems to be built on, and how these compare to other options available. Finally, this chapter will also lay out the research questions that this dissertation seeks to investigate and the methodology that will be used.

2.1 Speech-to-Text

Speech-to-text is the process of transcribing speech into written words. This is useful in many domains, and as noted by Jones (2022), it is not only transform- ing how people interact with digital content but also improving accessibility for people with disabilities in the digital world. Whether that be through automatic subtitling or speech assistants, many technologies rely on effective and accurate automatic speech recognition models.

Speech-to-text generally works by transforming audio into a spectrogram representation of the audio. This data can then be processed by an artificial neural network as a numeric series of data. The model that transforms this raw data into its textual representation, known as the acoustic model, is a foundational part of any speech-to-text system.

This process often makes use of the beam search algorithm. Similarly to the Viterbi algorithm, the beam search algorithm calculates the probability that a certain timeframe corresponds to each letter based on the output of the neural network and based on previous timesteps. The beam width is the variable that determines how far back the algorithm looks, the higher the beam width the further back it looks. Using a higher beam width will improve the overall performance of the algorithm but would use a significant amount of memory and computing power (Farhat, 2022).

Coqui and other systems often allow for an additional language model to be used to aid the beam search algorithm. This language model is trained on raw text and is used to improve the probabilities for certain characters following each other and helps to fix any misspelt words. The addition of a language model tends to improve the word error rate (WER) but generally does not improve the character error rate (CER) (see section 3.2.1 for the definition of these metrics). For example, Coqui (Coqui AI, 2022) uses a recursive neural network (RNN) for its acoustic model and allows for the addition of a KenLM language model (Heafield, 2011) to be used to aid the model during the beam search phase.

2.2 Phonology of Breton and Welsh

To understand the purpose of this project, it is important to understand some fundamental aspects of the phonology of both Breton and Welsh, and most importantly how they compare to the phonology of French and English. Both Breton and Welsh are Brythonic languages, having evolved out of Common Bry- thonic from the late fifth century and mid-sixth century (Willis, 2009). Despite the languages sharing a common origin, they have developed far enough apart that they are no longer mutually intelligible (Sims-Williams, 2015). One of the contributing reasons for this is the contact that Breton has had with French and Welsh has had with English. The syntax, morphology, and phonology of Breton have been influenced by French while Welsh has been influenced by En- glish. A noticeable example of this is that French and Breton are among the few languages in Europe that have nasal vowels, while English and Welsh both lack these phonemes (Sims-Williams, 2015).

2.2.1 Phonology of Breton

Breton has a large and wide-ranging phonemic inventory. According to Hemon and Everson (2011) it has 30 consonants, some of which are devoiced in certain circumstances. It also has at least 11 vowels, most of which can also be elongated when stressed. In contrast to French where only four vowels can be nasalised (ɑ, ɛ, ɔ, œ), all Breton vowels can be nasalised. In total that means that Breton has somewhere between 60 and 70 different phonemes in its phonemic inventory. This is a substantial amount, especially when compared to languages such as Welsh which has around 40 when only counting consonants and monophthongs (Cooper et al., 2019).

The phonology of Breton has been influenced a substantial amount by French over the years. There are many examples of this, most notably in the phonetics of the language and the shared phonemic inventory of the languages. While some dialects of Breton such as Leoneg pronounce “r” using the apical trill r, many dialects use the uvular trill ʁ instead (Hemon and Everson, 2011). This sound, also called the guttural r, is found in standard French as well. This is in contrast to English and Welsh which uses other types of rhotics like trills r, taps ɾ, and retroflex approximants ɻ. Note that taps ɾ and retroflex approximants ɻ

and found in parts of Tregerieg as well (Hemon and Everson, 2011).

Another notable feature of Breton is its nasalised vowels. As mentioned above, all vowels in Breton can be nasalised. This is not found in either English or Welsh, but it is found in French. While nasalised vowels are more restricted in French, it is a prevalent phonetic feature that they share that most other languages do not have.

Traditionally Breton tended to stress the penultimate syllable. This is some- thing that it shares with Welsh which is also stressed on the penultimate syl- lable. Despite this, there is evidence that the stress patterns have started to shift, especially in younger speakers (Kennard, 2021). These new stress pat- terns are influenced by French. This is yet another way in which French has been influencing the phonology of Breton over the centuries.

2.2.2 Phonology of Welsh

Welsh has a more limited phonemic inventory when compared to Breton. The number of consonants is comparable with Welsh having 29 consonants Cooper et al. (2019). According to Cooper et al. (2019), Welsh has up to 13 monoph- thongs and 13 diphthongs. Dialects of Welsh vary substantially between North and South Welsh. Despite this phonetically this only manifests itself in the vowel inventories of the different dialects with the consonant inventory being consistent across the different dialects.

2.3 Transfer learning in general

Transfer learning is a technique used in machine learning contexts to transfer knowledge gained from training one model to a different model where the input data or learning tasks differ. In the context of natural language processing, this could be using a model trained for one language to help train a model for a different language.

Very specifically, it is the process of taking knowledge gained from a source domain D_s learning task T_s and using that to improve the learning processes for another target domain D_t and T_t (Pan and Yang, 2009; Ruder, 2019).

Transfer learning is useful in many contexts, and has been used to create state-of-the-art results in different domains. For natural language processing, it has been used successfully by many people (Salimzianov, 2021; Tyers and Meyer, 2021; Bansal et al., 2018) to create effective models forlower-resourced languages by exploiting available data and models from languages with more available data such as English.

Pan and Yang (2009) differentiate between three types of transfer learning; transductive, unsupervised, and inductive transfer learning. The difference lies in the relationship between the source and target domains and learning tasks.

When the source domain and target domains are the same $D_s = D_t$, but the learning and target domain differ $D_s \neq D_t$, but the learning tasksare. When the source domains

the same $T_s = T_t$, this is called transductive transfer learning. Unsupervised transfer learning is when both the source and target domain and learning tasks differ $D_s \neq D_t$ and $T_s \quad T_t$.

In the context of natural language processing, the different domains and learning tasks are often different languages, and the difference often lies in whether you have labelled data for the target domain D_t (i.e the language that you are transferring the knowledge to). Ruder (2019) makes some further dis- tinctions between some subcategories of the types of transfer learning described by Pan and Yang (2009).

If we only have labelled for the source domain D_s (i.e the language we are transferring knowledge from), this would fall under transductive transfer learning since the source and target domains are not the same $D_s = D_t$. Since the learning tasks are targeting different languages this is a type of transfer learning called cross-lingual learning. If we have labelled data for the target domain D_t, this would fall under inductive transfer learning. If the transfer learning hap- pens in sequence, then this type of transfer learning is called sequential transfer learning.

This is a helpful distinction that Ruder (2019) makes, and makes it easier to distinguish between different types of transfer learning used in natural language processing. Many forms of transfer learning used in natural language processing, especially in speech-to-text, use a model that is trained on the source language D_s to achieve its to perform its learning task T_s. This model is then used as a basis to train a model for a different domain D_t to perform the same task. For example, for speech-to-text and English STT model can be trained and used as a basis for a Welsh model. This is an example of sequential transfer learning and is a common method of transfer learning for STT tasks.

2.4 Lower-resourced languages and transfer learn- ing

A lower-resourced language is generally defined as a language that has a lack of or limited availability of elements like a fixed orthography, presence in the digital space, and digital resources (Besacier et al., 2014). Digital resources such as online dictionaries, pronunciation dictionaries, corpora (both written and spoken), and so forth are vital to building digital tools and services for languages. Lower-resourced languages often lack many of these resources, which makes creating tools and services and enabling the language to thrive in the modern world difficult. According to Besacier et al. (2014) there are more than 6900 languages in the world and only a small number of these have sufficient resources available. Both Breton and Welsh are considered lower-resourced languages.

While lower-resourced languages do not have to be endangered or minor- ity languages, the reverse is often true meaning that minority and endangered languages tend to be lower-resourced Besacier et al. (2014). This creates a sig- nificant imbalance where languages that are more well-off are able to thrive and

be more used in the digital world leading to more available data, while languages that are endangered and need these resources are falling behind. This imbal- ance is further manifested in the fact that teams working with lower-resourced languages tend to be smaller and have fewer resources like GPUs available to work with.

This is where modern machine-learning techniques such as transfer learning become extremely important. By utilising available data from more well-off languages, effective models can be created for lower-resourced languages. In re- lation to speech-to-text, collecting speech data is often a tedious and expensive process, but by utilising data available for languages such as English we are able to create effective models using a significantly lower amount of data. With the advent of open-source corpora and resources such as Common Voice (Ardila et al., 2019), data and tools for minority languages have become democratised, which has enabled more data to be made available. Both of these elements together have had the effect of lowering the barrier of entry to technologies and have enabled technologies like speech-to-text to be developed for these lan- guages. These technologies, tools are services are vital to the digital presence of minority languages, and by enabling these technologies more minority languages are able to thrive in the modern world.

2.5 Transfer learning in relation to STT

Transfer learning has been used successfully in the past to create models for lower-resourced languages. Both Salimzianov (2021) and Tyers and Meyer (2021) showed that effective results can be achieved utilising transfer learning in speech-to-text models range of languages. Even though the resulting models are not as effective as state-of-the-art, they are still better than what could be achieved without the usage of transfer learning and they provide a valuable stepping stone in the process of enabling better models for these lower-resourced languages.

Tyers and Meyer (2021) also showed that careful selection of parameters is important to achieving good results and that just fine-tuning the parameters can result in a 5% to 15% decrease in character error rates (CERs). The same trend has been shown in other research as well. This highlights the importance of the parameters for each language and shows that consideration will have to be taken to ensure that the parameters are optimised for the language in question. Transfer learning has also been used in other natural language processing domains as well, such as speech-to-text translation, to great effect. Research conducted by Bansal et al. (2018) seems to indicate that the more training data the better, even using data that are unrelated to the target languages. Their Mboshi-to-French model performed better when using a model that was pre-trained in both English and French. Their approach using the French-only pre-trained model performed better than their English-only one despite having a significantly lower amount of data. This seems to indicate that the language combination has some effect on the results. Despite this, this has not been

properly explored for standalone speech-to-text models using transferlearning. Tyers and Meyer (2021) used the same pre-trained English model for all of their experiments and Salimzianov (2021) also used a single pre-trained English model.

The decision to use English as a base to build models is understandable, but it highlights an inherent bias towards using specific languages as a base without taking into account the suitability of that language in the context in which it is used. Rahimi et al. (2019) showed that the base language can have a considerable impact on the quality and performance of models when using direct transfer. This has not been explored in depth in relation to transfer learning for speech-to-text is the choice of language to use as a base, and it raises the question of whether the results that Tyers and Meyer (2021) and Salimzianov (2021) obtained can be improved by utilising a different set of source languages.

2.6 Common Voice

Common Voice is an open-source and crowd-sourced project that contains speech corpora for a wide range of different languages (Ardila et al., 2019). Due to the crowd-sourced nature of the project, the audio is not of high enough quality for certain speech-related tasks such as text-to-speech. However, Common Voiceis a valuable resource for training speech-related technologies such as speech-to- text.

For languages such as English, Spanish, Catalan, French, and others, there are enough data to train speech-to-text models from scratch. However, lan- guages like Breton and Irish only have nine and four hours of audio respectively. While this is a substantial amount of data, it likely is not enough to train an effective model from scratch. However, by utilising transfer learning, it might be possible to make some useful models for these languages despite them being under-resourced. Breton was one of the languages that Tyers and Meyer (2021) tested their system on and they did get good results, but not as good asmany of the other languages they tested.

2.7 Coqui STT

Coqui STT is an end-to-end speech-to-text framework developed by Coqui AI. It is an independent continuation of Mozilla's Deep Speech framework. Deep Speech is based on a recurrent neural network (RNN) which is trained to ingest spectrogram data (Hannun et al., 2014). This RNN is built using Tensorflow (Abadi et al., 2016). Coqui uses a modified version of this architecture[1]. Since Coqui is an end-to-end speech-to-text framework, it is trained on transcribed text rather than phonemes. This also means that it does not require a pronunci- ation dictionary or Grapheme-to-Phoneme (G2P) model and it makes the Coqui

[1]Information about the architecture of Coqui can be found at https://github.com/ coqui-ai/STT/blob/main/doc/Architecture.rst

framework language independent. This stands in contrast to earlier systems like Kaldi (Povey et al., 2011) and HTK (Young et al., 2002).

Coqui STT has some benefits over other similar systems like wav2vec (Baevski et al., 2020) in that it supports streaming, i.e transcription of audio on-the-fly as opposed to requiring a complete audio file. This usually results in worse over- all performance for the models. As mentioned by Salimzianov (2021), it also has far less demand for computing. Other than that, it also supports transfer learning and support for additional language models to be inserted. This means that it is possible to pre-train base models using Coqui, and then use these as a foundation for other models.

As mentioned earlier, Coqui STT was used by both Salimzianov (2021) and Tyers and Meyer (2021) to great effect. There are however some questions that they left open. They only used one English model for all of their experiments and did not use any other base language or any combination of languages like Bansal et al. (2018). This leaves the question: Do the base model and the language of the base model have an impact on the performance of the model or is any measurable difference simply down to the amount of training data? And if it does impact the overall performance, how significant is it?

If we look at languages that are closely related, like Breton and Welsh, there are still characteristics that differ between the languages. For example, the realisation of the letter "r" is often in the French-inspired standard version of Breton realised as /ʁ/. Welsh on the other hand does not use this sound, and neither does English. French on the other hand does use this sound. There is a likelihood that the model could learn to recognise this sound from the French dataset, and therefore any model that has been trained on the French base model would be better at recognising this sound than if it was trained on the English base model that had not come across the sound before.

This boils down to a question about quality versus quantity. Do the type of data and the phonology of the language of the base model affect the performance of the model? This is something that has not been investigated by the existing research and is something that could play a role in determining how to maximise the benefits gained from transfer learning. This leaves an opening in the existing literature and is something that this project seeks to investigate further.

2.8 Summary and Discussion

We have discussed the issues facing lower-resourced languages and why tech- nologies such as transfer learning can play a crucial role in enabling digital tools and services to be developed for these languages by exploiting available resources from other languages. Newer speech-to-text frameworks like Coqui have in-built functionality to do transfer learning, making it possible to create effective models for these languages without a vast amount of data.

It is also clear that there are some unanswered questions in the existing lit- erature in regard to how to maximise the effectiveness of transfer learning in a speech-to-text context. A methodology has been proposed that will enable

the dissertation to attempt to answer some of these questions. This will hope- fully enable more effective models to be trained, meaning that lower-resourced languages have the opportunity to thrive in a digital world and enable better accessibility services to be developed.

Chapter 3

Research Question, Hypotheses, and Methodology

This chapter will lay out the main research questions that this dissertation will attempt to investigate. In addition, a set of hypotheses will be formalised so that these can be evaluated during the analysis. The chapter will also go into more detail about the methodology that the dissertation will use to try to answer the research question. Finally, the chapter will discuss what metrics are available to be used to evaluate the performance of the acoustic models, how they are defined, and which ones will be used for this dissertation.

3.1 Research questions and hypotheses

Based on the review of existing literature, it is clear that there are some ques- tions in relation to transfer learning and speech-to-text that are not sufficiently answered by this existing literature, such as what impact the choice of base lan- guage has on the effectiveness of the transfer learning. As we have seen, there are examples in other domains where it has been shown to have a substantial ef- fect. Therefore, the main hypothesis of this dissertation is to investigate whether the choice of base language does in any way impact the overall performance of models.

Hypothesis 1: *The choice of base language does impact the overall perfor- mance of the models.*

If this hypothesis holds true, we would expect models that are trained using languages that have a high degree of "compatibleness" to perform better than models that are trained using languages that have a lower degree of "compati- bleness". What "compatibleness" means is very abstract, but this dissertation

will attempt to quantify this relationship and use these quantitative metrics to evaluate the merit of the hypothesis.

Based on this hypothesis, its corresponding null hypothesis can be defined as such:

Hypothesis 0: *The choice of base language does not impact the overall perfor- mance of the models.*

It might very well be that the dissertation will show that there is no merit to the hypothesis and that other metrics such as the amount of training data are of much greater importance. This dissertation will also attempt to take into account these other metrics and attempt to figure out whether this is the case.

3.2 Methodology

To investigate the basis of the hypothesis that the base-model language can impact the quality of the final model, an analysis of the distribution of phonemes and the phoneme inventory of the different languages and how these compare to the other languages will be carried out.

In total six models will be trained based on three base models. One English, one French, and one French model will be created using transfer learning from English. An English model already exists and is available from Coqui's repos- itory. As such, there is no reason to train a new one from scratch. However, the two French models will have to be trained. Then one Welsh and one Breton model will be trained based on all of the three base models. This means that we'll have six data points for our evaluation.

The data for this training will be taken from the Common Voice speech corpus. This data has to be cleaned and pre-processed before it can be utilised by Coqui STT as training data for our models. There is a substantial amount of data available for both English and French, with 2,224 hours of validated speech data for English and 848 hours for French. Welsh has substantially less with 117 hours available. Breton however, only has 9 hours of validated data (Ardila et al., 2019).

The parameters that are chosen for the training of these models are impor- tant. Carefully selecting the parameters could improve the overall performance of the speech-to-text models with between 5 to 15% (Tyers and Meyer, 2021). Therefore some preliminary testing needs to take place in an attempt to op- timise these parameters. The project will aim to make smaller models using a subset of the training data, and then evaluate the validation loss curves to determine the parameters that are likely to yield the best results.

Using these parameters, full models will be trained using the full available dataset. Given the amount of data, this training is likely to take some time.

In addition to this, an analysis will be conducted of the data, which will aim to extract several metrics about the dataset. This includes the overlap of phonemes between two languages, the Euclidean distance between the rela- tive frequencies of phonemes between two languages, the number of phonemes

missing and so forth. The aim of this is to see whether there is a correlation be- tween the character error rates of the final models and the interaction between the base language and the target language. The main metric that will be used is Pearson's correlation coefficient (r) (Pearson, 1895).

3.2.1 Performance metrics for acoustic models

There are two common ways of measuring the performance of speech-to-text models. That is the word-error-rate (WER) and character-error-rate (CER). Both use the Levenshtein distance (Levenshtein et al., 1966) as a way of mea- suring the distance between the output of the speech-to-text model and the ground truth. The Levenshtein distance is defined as the minimum amount of operations such as additions, substitutions, and removals that have to occur to transform one string into another. While there are many ways of implementing the Levenshtein distance algorithm, a recursive implementation that computes the distance between string a and b can be seen in equation 3.1:

$$\text{lev}(a, b) = \begin{cases} |b| & \text{if } |a| = 0, \\ |a| & \text{if } |b| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise.} \end{cases} \tag{3.1}$$

There are a couple of things to note about this implementation. The recur- sive implementation of the Levenshtein distance as seen in 3.1 will make the same recursive call multiple times. Most implementations will store intermedi- ate results to mitigate that issue. Strings are also zero-indexed as they would be in computer implementations. The tail function as seen in equation 3.1 is defined as in equation 3.2:

$$\text{tail}(a) = \begin{cases} [\,] & \text{if } |a| = 0 \\ a[1, |a|] & \text{otherwise.} \end{cases} \tag{3.2}$$

The difference between the word error rate and the character error rate is whether they count the number of words that need to be corrected or the number of characters. The Levenshtein distance is in both cases divided by the length of the ground truth so that the metric becomes independent of the length of the string. The definition of the character error rate (CER) metric can be seen in equation 3.3. The definition of the word error rate is exactly the same as the character error rate, but instead of calculating the Levenshtein distance using the characters of the strings, the words of the strings are used. Since any misspellings will cause the entire word to be classified as wrong, the word error rate tends to be higher than the character error rate.

$$\text{CER}(target, output) = \frac{\text{lev}(target, output)}{|target|} \quad (3.3)$$

When measuring the performance of only an acoustic model, the most useful metric is the character error rate. Since no language model is being used, misspellings are not being corrected. This means that the word error rate is very closely correlated with the character-error rate and it simply measures the rate at which a word has no misspellings. The lower the character error rate the higher the chance that a word will be properly spelt. As such, using the character error rate itself gives in this case a more accurate picture of the acoustic models' ability to detect and correctly classify the sounds.

The character error rate is not a perfect way of measuring the performance of models. Depending on the orthography of the language, there might be a significant distance between a character and a phoneme. This is especially true for languages such as English and French which has an orthography that is less phonetic than languages such as Breton and Welsh.

A phoneme error rate of the models could potentially provide a better way of quantifying the performance of the models. This is because it more closely represents the performance of the model in terms of phonology and removes any effect the orthography of the language might have on the results. For this project, it could be more helpful to train and evaluate models using phonemes rather than text. There are some issues with that approach, however. While systems like Kaldi and HTK use phonemes when transcribing audio, Coqui does not. Coqui is, as mentioned in section 2.7, an end-to-end speech-to-text framework meaning it does not rely on pronunciation dictionaries and it outputs characters as opposed to phonemes. This makes measuring thephoneme-error- rate could difficult.

Another issue is that there is not enough data to accomplish this. For example, there is no freely available pronunciation dictionary for Breton, and very few corpora, if any at all, with phonetic transcriptions exist for languages such as Breton and Welsh. This makes it incredibly difficult to properly train and evaluate the models.

Due to all of the aforementioned reasons, it was decided that using the character error rate as the main metric was the best option.

3.3 Summary and Discussion

In this brief chapter, we have defined the main research questions of the dis- sertation and formulated some hypotheses based on them. The methodology describing how the dissertation is going to attempt to answer these questions was also discussed.

We have also had a look at the different evaluation metrics for acoustic models and determined that the character error rate is the best metric that can be reasonably used given the resources that are available for the project. A formal definition for the character error rate was also defined.

Chapter 4

Preparing and Analysing the Training Data

This chapter describes the work undertaken before any of the acoustic models were trained. That mainly includes the preprocessing and analysis of the Com- mon Voice datasets. There are two reasons why this was carried out. Firstly, to create efficient models, it is beneficial to know and understand the underlying data, any issues it might have, or any particularities that need to be accounted for. Secondly, in order to perform the analysis later, we need to extract some metrics about certain characteristics of the languages and the training data. This chapter explains the work that was completed in relation to these two points and discusses the implications of the findings and the consequences that these might have on the overall results of this experiment.

4.1 Amount of data in the Common Voice datasets

Common Voice 10 was released during the project. This release contained new data for all four languages. The amount of data for Welsh and Breton can be seen in figure 4.1. The data increase from Common Voice 9 is not very significant, especially for Breton and Welsh, and the rate of increase seems to have been slowing down considerably in the past few releases. This is rather unfortunate and highlights that there is still substantial work to be done to facilitate and motivate people to contribute to projects such as Common Voice. This is especially true for minority languages.

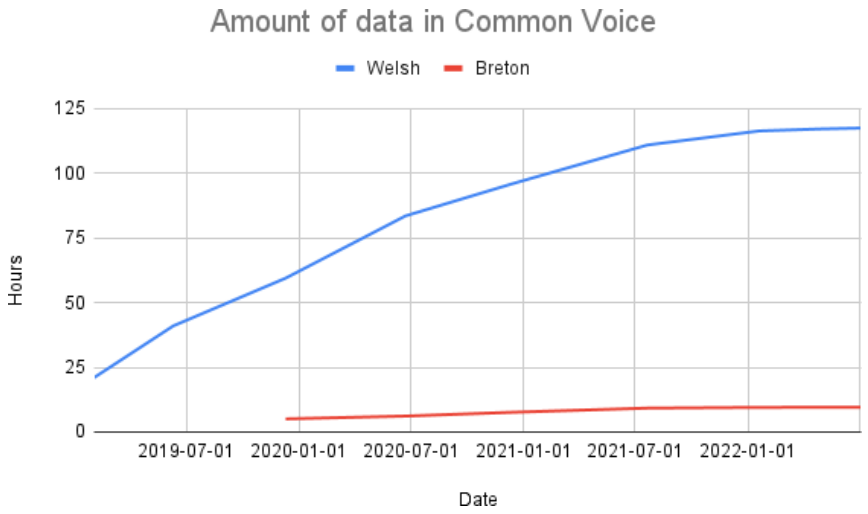

Figure 4.1: The amount of data in the Common Voice dataset for Breton and Welsh. Note that some of the early data for Breton is missing.

While it is difficult to see in figure 4.1, figure 4.2 makes it more obvious that this trend is not only true for Welsh but is also affecting Breton.

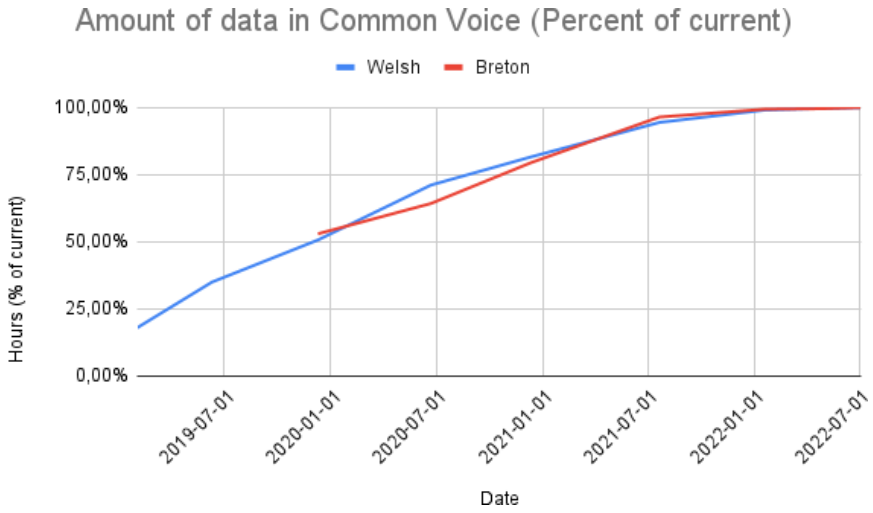

Figure 4.2: The amount of data in the Common Voice dataset for Breton and Welsh over time compared to the amount of data in Common Voice 10.

4.2 Unique sentences in the Common Voice datasets

All Common Voice datasets come with a set of predefined splits. This includes a training set, a testing set, and a validation set (also called a development set). These datasets only use the validated data, which is data that have been reviewed and validated by the Common Voice community.

As noted by Jones (2022), Mozilla's predefined datasets only use a single sentence across all of its sets. This means that for languages that have a high rate of duplicated sentences, the amount of usable data diminishes greatly. To investigate the severity of this problem, a BASH script[1] was created toextract the number and percentage of unique sentences in the validated set. The results that this script produced can be seen in table 4.1.

Language	Total	Unique	%
Breton	11169	6875	61.55
English	1589009	954095	60.4
French	625587	491052	78.49
Welsh	87295	18004	20.62

Table 4.1: Overview of the number of unique sentences in the validated.tsvfile for each language in Common Voice 10.

As can be seen in table 4.1, the Welsh dataset is unique in having a very high rate of duplication. This means that the effective dataset is significantly lowered. To compensate for this, new custom splits will have to be made. By creating custom splits more of the validated data can be used, and hopefully, will lead to improved models.

4.3 Analysis of phoneme distribution in thetrain- ing data

To investigate the basis of the hypothesis that certain shared characteristics of base and target languages impact the overall performance of the models, an analysis of the phoneme distribution in the training data had to be under- taken. This was achieved by getting a pronunciation dictionary for each of the languages, converting the sentences into their International Phonetic Alpha- bet (IPA) (International Phonetic Association et al., 1999) representation, and then looking at the distribution of IPA symbols and how they differed between languages.

[1]The script can be found at https://gitlab.com/prvInSpace/master-dissertation/-/blob/master/data/common_voice/unique.bash

4.3.1 Pronunciation Dictionaries

A set of pronunciation dictionaries were required for this experiment. The reason is that we wanted to analyse the phonemes used within the Common Voice datasets. In order to perform this analysis, sentences in Common Voice had to be converted into IPA so that presence of and the number of occurrences for different phonemes could be documented for each language. Pronunciation dictionaries were used to convert words into their IPArepresentation.

There are some drawbacks to this approach. Firstly, many pronunciation dictionaries might be incomplete. This leads to a loss of information because we cannot analyse those words. This is especially a problem for languages such as Breton and Welsh that has initial letter mutations because the radical forms (i.e the unmutated forms) of the word might be present but sometimes the mutated forms are missing. In certain circumstances, this could be rectified by attempting to de-mutate a word and automatically changing the initial letter sound. One issue, however, is that sometimes this change in sound have a knock-on effect that affects the realisation of the following sounds.

Another issue is that of homographs. There might be words in different languages that are spelt the same way but pronounced differently. The simplis- tic approach taken here will have a hard time distinguishing between the two homographs and will be forced to choose one over the other. It should be noted that homographs are not too common. While the results are affected by this issue, it should not change the results enough to invalidate them.

Finally, not all languages have publicly available digital pronunciation dic-tionaries. For languages like English and French, there are comprehensive and publicly available pronunciation dictionaries. The same can not be said for some of the other languages, especially Breton which had no publicly available pro- nunciation dictionary. How this issue was overcome will be discussed later in this section.

Publicly available pronunciation dictionaries used

For Welsh, the Bangor Pronunciation Dictionary (Jones and Cooper, 2021) was used. This repository also contains a pre-processed copy of the CMU English Pronunciation Dictionary by Weide et al. (2015). While the CMU Pronuncia- tion Dictionary uses the ARPAbet format (Barnett, 1975) to transcribe words, Jones and Cooper (2021) has a version where the ARPAbet entries have been translated to IPA. Given that it was in the same format as the Welsh pronun- ciation dictionary, this dictionary was used for English.

Breton Pronunciation Dictionary

There seems to be a lack of freely available pronunciation dictionaries online, hence, a pronunciation dictionary for Breton had to be created. To achieve this, a Python script was created to scrape the Breton version of Wiktionary, called Wikeriadur (Wikeriadur contributors, 2022). While this provided a starting ba- sis, it was clear that the pronunciation dictionary was not perfect and contained

many mistakes that probably happened as part of the scraping. Therefore the dictionary had to be checked, cleaned up, fixed, and verified. With the help of some native Breton speakers, this work was carried out by Vangberg et al. (2022) and the work was made available online for others to benefitfrom.

4.3.2 Methodology

A Python program[2] was developed to carry out this task. This program has to achieve a couple of things: extract a list of sentences from Common Voice, convert these sentences into their IPA equivalent, and then count the occurrences of the different phonemes in the different languages.

The Common Voice datasets come with a set of tab-separated values (TSV) files that contain information about all of the audio clips. The sentences for the different languages could then be extracted from the validated set which contains a list of all of the audio files that have been validated by the community. Since there might be several recordings of the same sentences, the list were then filtered for duplicates so that each sentence only get processed once.

[2]Can be accessed at https://gitlab.com/prvInSpace/master-dissertation/-/blob/master/scripts/phoneme_distribution.py

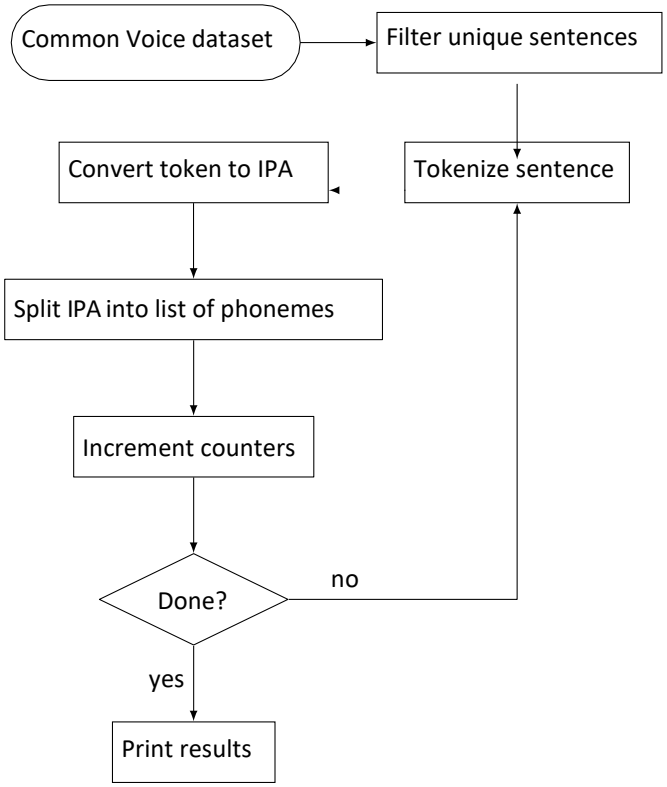

Figure 4.3: Flowchart of the process of extracting phonetic data from the train- ing data.

As explained in the previous sections, these sentences had to be converted into their IPA representation. This conversion is done by tokenizing the text and then looking up the tokens in the pronunciation dictionary for the language being processed. These tokens can then be further split so that each phoneme is separated from the ones around it, but keeping any markers such as prolongation and similar. When that has been done, it is just a case of counting the number of occurrences for each phoneme in the dataset. This process allows us to analyse what phonemes are present and their relative frequencies of phonemes. These metrics can then be transformed into usable metrics that can be used for our analysis.

4.3.3 Results

Before we look at the results, let us look at how well the script was able to convert tokens to IPA and how many phonemes it found for each language. A summary of this can be seen in table 4.2.

Language	% Converted	Phonemes found
Breton	72.77%	68
English	96.36%	38
French	87.73%	82
Welsh	98.39%	43

Table 4.2: Overview of how many phonemes were found and how many tokens were successfully converted to IPA.

As can be seen in table 4.2, the script was able to successfully convert most tokens for most languages. French and Breton have a significantly lower con- version rate than both English and Welsh. For Breton, the main reason for this is that the pronunciation dictionary does not contain any mutated form of the different tokens. Therefore, only the radical (unmutated) forms got successfully converted.

The amount of phonemes extracted for each language is also interesting when we compare this to the number of phonemes that we expected. According to Cooper et al. (2019), there are 29 consonants, up to 13 monophthongs, and up to 13 diphthongs in Welsh. Since the script does not recognise diphthongs, this would mean that we would expect 42 phonemes in Welsh. That is very close to the 43 number that we found.

The same is true for Breton. According to Hemon and Everson (2011), there are 30 consonants and 11 vowels where most of which can be elongated and nasalised. That would mean that Breton has around 63 phonemes. The number we got, 68, is slightly higher than this, but this number includes allversions of "r" found in Breton in addition to sounds that are likely from borrowed words. English is roughly aligned with expectations as well. According to Yavas (2020), there are 24 consonants and 12 monophthongs in American English. This means that we should expect around 36 phonemes. This aligns pretty well with the results that we have found and the two additional phonemes are elongated vowels.

The number of phonemes in French is likely inflated. According to Hannahs (2007), there are 21 consonants and 11 monopthongs in French. Note, that this does not include elongated forms, the four nasalised vowels, and schwa ə.
If we assume that every vowel can be elongated, that means that there are a total of 49 different phonemes in French. This is substantially lower than the 82 that our script produced. The reason for this is that the script found asubstantial amount of phonemes that are not present in the list in Hannahs (2007). For example, it includes all rhotic consonant variations and allophones of the phonemes in Hannahs (2007). This is likely due to the pronunciation dictionary containing a much narrower transcription of the words than some of the other dictionaries.

How this will impact the results is hard to determine. The number of phonemes for the other languages is mostly aligned with expectations. For

this project, it was determined to stick with the results returned by the script on the grounds that if the sounds are in the datasets then they should be in- cluded. Any future research should take this into account, however, and should probably aim to normalise the transcriptions so that the different languages are transcribed approximately with the same broadness.

An overview of all of the relative frequencies of the phonemes can be seen in figure 4.4 and 4.5. As expected, the most commonly shared phonemes have quite similar relative frequencies. Since many unstressed vowels in English get turned into schwa ə, it is not surprising that this phoneme is very common in English.

However, this data is not very useful on its own, but there are some metrics that can be extracted from it that we can use for our analysis. This is what we will discuss next.

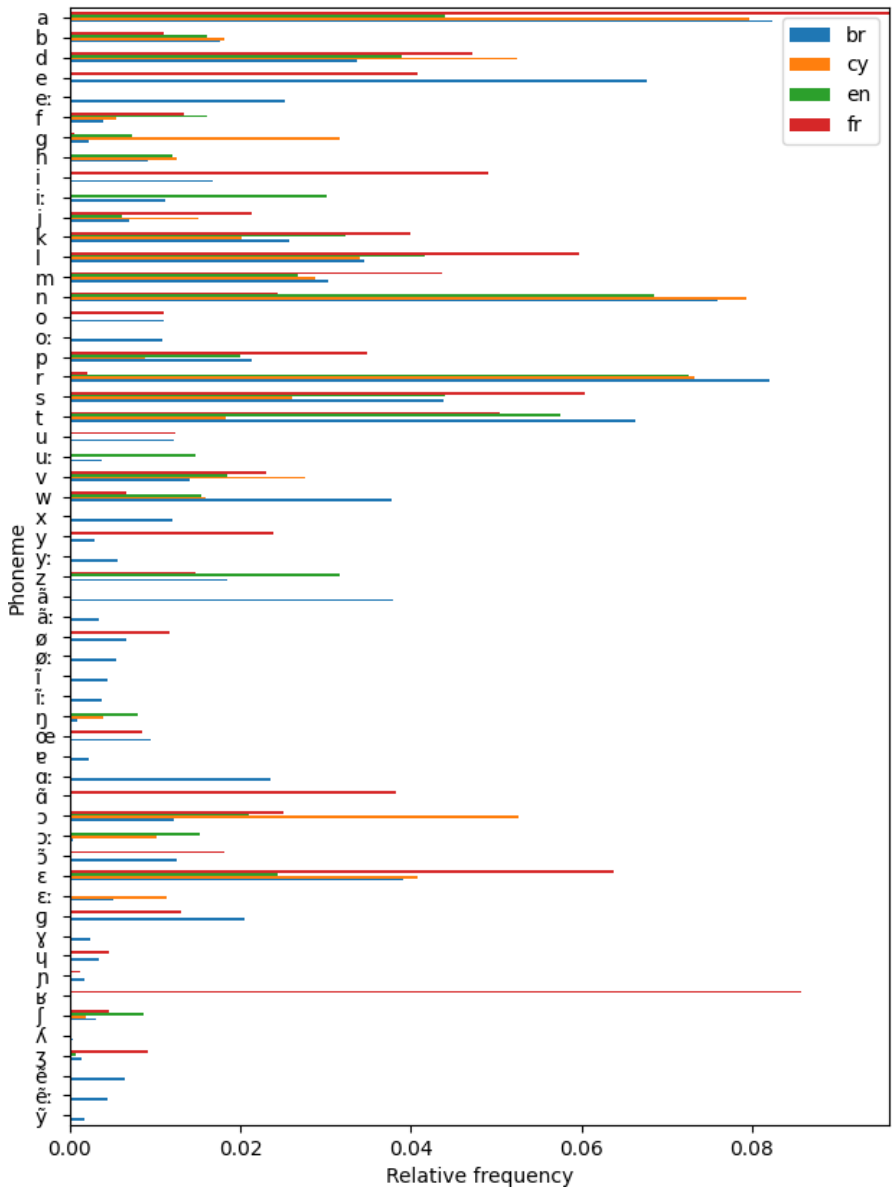

Figure 4.4: Overview of the relative frequency of phonemes in the different languages (part 1).

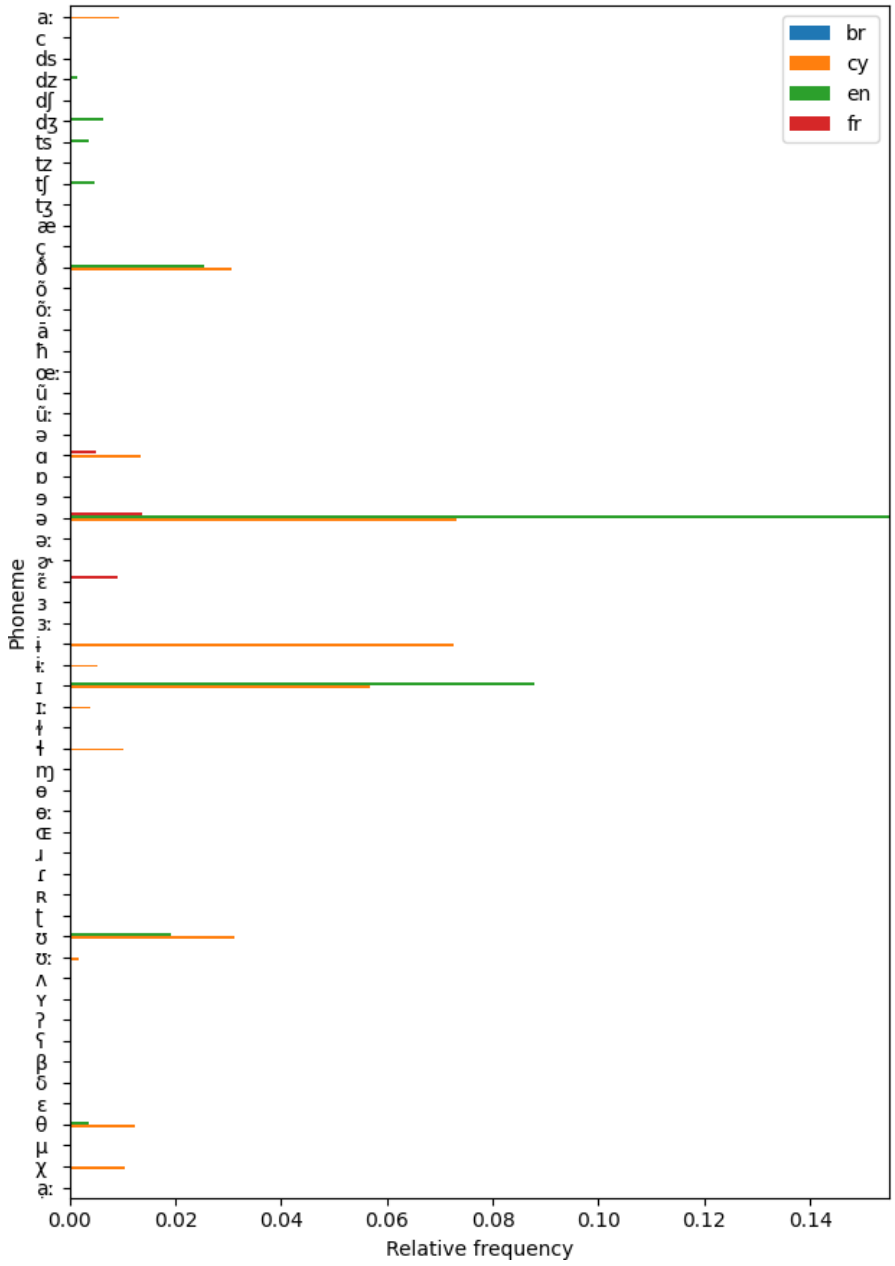

Figure 4.5: Overview of the relative frequency of phonemes in the different languages (part 2).

4.3.4 Extracting metrics for analysing the hypothesis

There are several metrics that were extracted from this data. These metrics were then used to see if there was any correlation between them and the character error rates of the models. The three metrics that were considered of interest were: the percentage of phonemes in the target language that is also present in the base language; the Euclidean distance between the relative frequencies of phonemes between the two languages; and the number of phonemes in the target language not present in the base language.

Phonemes in the target language also present in the base language

One of the core ideas of hypothesis 1 (see section 3.1) is that when training a language based on a language that has a low overlap in phonemes, this means several new unseen patterns would have to be learnt during training. One way of quantifying this is to look at the percentage of phonemes in the target language that is also present in the base language. This can be shortened to just the phoneme overlap between the two languages.

If B is defined as the set of phonemes in the base language and T is defined as the set of phonemes in the target language, the phoneme overlap (PO for short) between the two languages can be calculated as in equation 4.1:

$$\text{PO}(B, T) = \frac{|B|}{|B \cap T|}. \tag{4.1}$$

If hypothesis 1 is true, we should expect a negative correlation between the overlap of phonemes between the two languages and the character error rate, meaning the higher the overlap, the lower the character error rate.

Phonemes in the target language that are not present in the base language

As mentioned in the previous section, the difference between the phoneme inventory of the target language and base language is very much at the core of hypothesis 1. Another way to quantify this difference other than the phoneme overlap is the absolute number of phonemes in the target language that is not present in the base language. As with the phoneme overlap, if B is defined as the set of phonemes in the base language and T is defined as the set of phonemes in the target language, a number of phonemes that needs to be learned (PNP (Phonemes Not Present) for short) can be calculated as seen in equation4.2:

$$\text{PNP}(B, T) = |B - T|. \tag{4.2}$$

If hypothesis 1 is true, then these would be phonemes that the model would have to learn the patterns of during training. We should expect a positive correlation between this metric and the character error rates of the models, meaning the more phonemes have to be learned the higher the character error rate.

Euclidean Distance

One interesting aspect when comparing two different languages is how often a phoneme occurs in one language compared to another. If hypothesis 1 is true, it could be reasoned that if a sound is only rarely seen in one language but very frequently in another this might affect the quality of the finished model. In order to compare the relative frequencies of phonemes between two languages, the Euclidean distance was utilised.

Since the experiment above returned the absolute number of occurrences, it is possible to calculate the relative frequency of each phoneme in the language by dividing the number of occurrences *n* by the total amount of occurrences N as shown in equation 4.3:

$$f_i = \frac{n_i}{N}. \tag{4.3}$$

To calculate the Euclidean distance between the two languages a list of all phonemes present in either language is required. If $B_{|p \quad f|}$ represents the dictionary of phonemes in the base language and their associated relative frequencies in the base language and $T_{|p \quad f|}$ represents the dictionary of phonemes in the target language and their associated relative frequencies, the list of all phonemes in either language P can be derived by taking the union of the two sets of phonemes. This operation can be seen in equation 4.4:

$$P = B_p \cup T_p. \tag{4.4}$$

The original dictionaries have to be modified slightly as there is now the possibility that there are phonemes in P that are not present in the original dictionaries. This can be resolved by inserting 0 where required as shown in equation 4.5:

$$\text{For } p \in P: \quad X(p) = \begin{cases} X(p) \text{ if } p \in X \\ 0 \quad \text{if } p \notin X. \end{cases} \tag{4.5}$$

When that has been done, the lists should be of equal length and the Eu- clidean distance then be calculated as shown in equation 4.6. This is based on the higher-order Euclidean distance formula as formulated by Tabak (2014):

$$d(B, T) = \sqrt{\sum_{p \in P} (B(p) - T(p))^2}. \tag{4.6}$$

The Euclidean distance will then return a number where 0 represents a perfect match and the higher the number, the larger the difference between the two lists.

This was implemented in Python[3] using sets to get a list of all phonemes in either of the two languages as explained in equation 4.4. List comprehensions

[3]For implementation, see the function "calculate_euclidean_distance" in the file located at https://gitlab.com/prvInSpce/master-dissertation/-/blob/master/scripts/ phoneme_distribution.py

were then used to add any missing data as outlined in equation 4.5. Then the NumPy library (Harris et al., 2020) was used to calculate the Euclidean between these two lists because it contains in-built functions for doing this.

4.3.5 Overview of extracted metrics

All of the metrics discussed in section 4.3.4 were gathered for each possible model pair. The full overview of metrics can be seen in table 4.3.

The columns in table 4.3 are explained as follows. The target language is the language that we are trying to train a model for. ISO 639-1 two-letter codes are used to represent each language. “br” is used to represent Breton, “cy” means Welsh, “en” means English, and “fr” means French (International Organization for Standardization, 2002). The base column shows the language of the base model and follows the same naming convention as the target column. The % of phon. in base column shows the % of phonemes present in the target language also present in the base language. The Distance column shows the Euclidean distance between the target and base language. The final column shows the absolute number of phonemes in the target language not present in the base language.

Target	Base	% of phon. in base	Distance	Missing phon.
br	cy	39.71	0.180	41
br	en	39.71	0.211	41
br	fr	75.00	0.168	17
cy	br	69.23	0.180	12
cy	en	74.36	0.149	10
cy	fr	79.49	0.212	8
en	br	87.10	0.211	4
en	cy	93.55	0.149	2
en	fr	96.77	0.239	1
fr	br	68.92	0.168	23
fr	cy	41.89	0.212	43
fr	en	40.54	0.239	44

Table 4.3: Overview of the metrics extracted for each model pair

Some interesting observations can be made when looking at this data. As expected, we can see that Breton has both a significantly higher phoneme overlap as well as a lower Euclidean distance between it and French than for any other language. This seems to underpin the idea that French and Breton have signifi- cantly influenced each other over the last millennium. Interestingly enough, the language that is the second closest to Breton in terms of Euclidean distance is Welsh. Given that these two languages are closely related this might not be a surprise, but it is interesting to see it manifested in the results.

In terms of Welsh, the results are more ambiguous. While English has the lowest Euclidean distance to Welsh, it is French that has the highest phoneme overlap. This might seem surprising, but there are some ways of explaining this. If we look at the total amount of phonemes present in each language, we can see that French and Breton have significantly more phonemes than both English and Welsh. Due to this, the likelihood of French containing any of the phonemes in Welsh is quite high, though some are not present.

If we are looking at these results and comparing them to hypothesis 1 that certain aspects of the base language have an impact on the final model, there are some predictions that we can make. If hypothesis 1 is true, based on the phoneme overlap, Euclidean distance, and the number of missing phonemes, we would expect that for Breton there will be a benefit in training the model using a French base model. For Welsh, on the other hand, it is a bit more ambiguous. If the most important metric is the number of unseen phonemes, we would expect it to be beneficial for the Welsh models to be trained on a French base model. However, if it is the Euclidean distance that is the most important metric, we would expect English to be the best language to use as a base. All of that being said, since the differences between English and French are so low, especially in terms of unseen phonemes, it might very well be that any benefit is negligible or none at all.

4.4 Summary and discussion

In this section, we explained the work that was undertaken to analyse and pre- process the Common Voice datasets before the training of the models. Most importantly, we devised a method to extract phonemic information about the dataset for the different languages and defined set metrics that would be used in our analysis. It is interesting to see how close these metrics were to expectations and especially how close the number of phonemes was to what existing literature suggested it would be.

Despite this, there are some caveats. Firstly, French seems to have been narrowly transcribed to a greater extent than the other languages. This might impact the robustness of any findings that the analysis produces. It is clear what this negatively impacts the results and it is also clear that for metrics like the Euclidean distance, it should not affect it at all.

It was also discovered, in line with other contemporary research, that the predefined data splits in the Common Voice datasets provide a low utilisation of the available data. This is especially true for Welsh. Therefore, we can conclude that it will be beneficial to create custom splits that have higher utilisation of the available data. These custom splits might also benefit the models on the whole leading to better models.

Chapter 5

Training and Evaluating the new Models

This chapter will explain the work that was undertaken to create a language-independent training environment and the steps that were taken to train allof the models. It then describes any issues that were encountered and how these were overcome. The chapter also goes into detail about what the researchthat was carried out in relation to the dataset splits, and the effect of the different methods. Finally, the chapter will explain how the performance was analysed and how these results compare to the state-of-the-art.

As described in section 3.2, a total of eight models were trained of which six made up the data points for the evaluation. For each one of the three base models, a Welsh model and a Breton model were created.

5.1 Creating the training environment

As described in the methodology section (see section 3.2), Coqui STT was used to create the acoustic models for this experiment. Specifically, the Docker image for Coqui STT version 1.3.0 was used. Docker (Merkel, 2014) is a tool that allows you to create predefined images or environments and run several instances of that image. These instances are called containers as they function as contained units completely separate from each other. This allows code to run in a protected environment and makes it easier to ensure that dependencies are set up properly allowing for easier to run the application across several platforms. In addition to this, a set of scripts were created and added to the Docker image so that they could be used inside the training container.

5.1.1 Folder structure within the container

In an effort to maximise the reusability of the code, the folder structure within the Docker container was designed to be as language-independent as possible.

This means that the base language and target language can be specified at runtime and utilise the same Docker image.

There are three folders that are present on every Docker image. /data, /base, and /export. The /datafolder contains the data for the target language such as training files. For the purposes of this experiment, /data refers to the Common Voice directory for the target language. The /base folder contains the base model that is used for transfer learning. This is a folder containing pre-trained checkpoints for another language such as French or English. Finally, the /export folder is where all of the output files go. For during the training process this will be the intermediate checkpoints, but also the exported models and logs.

These folders are linked as Docker volumes on runtime as can be seen in code snippet 5.1. This code enables a Docker container to access files outside of the container as if it was inside the container itself. It also starts a container that can then be used to train a Coqui model. From the view of the container, the /data and /basefolders are simply folders containing data, but from the outside, we can manipulate which data sources we use. This means that the Docker image is language-independent and can be reused for all of the experiments.

Snippet 5.1: Makefile

```
1 train: .DOCKER_IMAGE
2         nvidia-docker run $(DOCKER_PARAMS) \
3                 -v ${PWD}/cv/$(LANG)/:/data/ \
4                 -v ${PWD}/base/$(BASE)/:/base/ \
5                 -v ${PWD}/models/$(LANG)/$(BASE):/export/ \
6                 --name=${USER}-stt-train-$(BASE)-$(LANG) $(IMAGE_NAME)
```

5.1.2 Training the models

Coqui's main training module is called coqui_stt_training.train. This is a Python script that takes a series of parameters that impact the training in different ways. In this section, we will discuss the most important ones, their impact, and what values were used. The call to coqui_stt_training.train can be seen in snippet 5.2 and the line number in the snippet is provided in parentheses after the parameters where applicable.

As mentioned in section 2.7, Coqui supports transfer learning. This is achieved by specifying a checkpoint to load from using the --load_checkpoint_dirflag (line 8) and specifying the number of source lay- ers to drop using the --drop_source_layers flag (line 11). Using theinternal folder structure as outlined in section 5.1.1, the model for the base language will always be located in the /base folder. Hence, --load_checkpoint_dir was always be set to /base. To avoid overriding the existing base model, the checkpoints were saved to /export/checkpoints. This was specified using the save_checkpoint_dir parameter (line 9). The --drop_source_layersflag is used to specify how many of the layers in the base model to drop. For this ex-

periment, it was decided to drop two layers as this seemed to produce effective models. Something to improve upon in the future would be to further investi- gate the effect of dropping different amounts of layers so that this feature can be utilised as effectively as possible.

Both a training set and a development set were used for each of the models. These are specified using the --train_files (line 4) and --dev_file(line 3) flags respectively. The development set was used to ensure that the models did not overfit against the training data. This will hopefully have ensured that the models are more general-purpose and should be more adaptable to new data.

Another flag that was used during training was --reduce_lr_on_plateau (line 13). This flag is used to reduce the learning rate when the training script detects that the training has plateaued. It is likely that this flag this not impact the overall results or improve the models in any measurable way. The reason for this is that in most cases, the model was still improving when the results for the development set started getting worse, meaning that training was stopped before the training had the chance to plateau.

Snippet 5.2: train.bash

```
1 python3 -m coqui_stt_training.train \
2       --load_cudnn true \
3       --dev_file /data/clips/prv_dev.csv \
4       --train_files /data/clips/prv_train.csv \
5       --train_batch_size 64 \
6       --dev_batch_size 64 \
7       --alphabet_config_path /data/alphabet.txt \
8       --load_checkpoint_dir /base/ \
9       --save_checkpoint_dir /export/checkpoints \
10      --use_allow_growth true \
11      --drop_source_layers 2 \
12      --max_to_keep 2 \
13      --reduce_lr_on_plateau true \
14      --epochs 150
```

A BASH script was created to make it easier to call the Python script and to clear out the /export folder. This script was called train.bash[1] and a code snippet showing the the call to coqui_stt_training.train can be seen in snippet 5.2. In addition to the previously mentioned parameters, a couple of extra parameters were provided. The first one is --load-cudnn (line 2). This flag is required for Coqui to be able to load the English model. The second one is
--use_allow_growth (line 10). This flag allows Tensorflow to grow the amount of memory it is using. The third one is --max_to_keep(line 12). This parameter allows the user to specify how many checkpoints to store. This was set to 2. Finally, a batch size of 64 was used for both the test set and the development set using the --train_batch_size (line 5) (line and --dev_batch-size (line

[1]The full file can be found at https://gitlab.com/prvInSpace/master-dissertation/-/blob/master/train/scripts/train.bash

6). This allows the model to process 64 files at a time instead of only 1. This speeds up the processing time.

5.1.3 Extracting loss values from the training environment

In order to create graphs documenting the loss values during the training, a small AWK program[2] (Aho et al., 1979) was devised to extract the loss values for the validation stage for each epoch. While training the model, Coquiprints updates to the terminal. By processing this text we can extract the loss values that are printed. While other methods were possible, like wrapping the Python training program or modifying the Coqui source code, given that Docker has a feature to access the logs of any given container, it was easier to just take these logs and filter out the wanted values using AWK.

The way this was done was quite straightforward, but there were a couple of issues that had to be resolved. Firstly, the output uses carriage returns to override the previous output. Therefore, the final loss value is at the end of some very long lines as opposed to being at the same position on every line. However, it was possible to simply loop over every element of the line and check whether it matches the format of a loss value and store if it does.

The other problem is that before starting the actual training process Coqui runs a dummy training process to ensure that it has the required memory and GPU capacity to run the actual training. The format of this is the same as any other epoch, and these values had to be filtered out. This was achieved by setting a flag once the dummy process was finished and only printing values if that flag was set.

5.2 Creating the dataset splits

As noted in section 4.2, the Common Voice dataset had a low rate of unique sentences. Since Common Voice's predefined splits only use one recording for each individual sentence, that results in the predefined splits having a very poor utilisation of the available data. Due to this, it was decided that it would be beneficial to create custom training, testing, and validation splits.

There is one issue to consider when splitting the data, and that is the issue of testing on seen data. However, in the context of speech corpora like Common Voice, what is "seen data"? This depends on whether we define utterances or sentences as unique data points, and whether training and testing on different utterances of the same sentence are considered testing on seen data.

To test whether splitting on an utterance level gave the models an unfair advantage, it was decided to test this by training both a Breton and a Welsh model using both methods.

[2]The AWK script can be found at https://gitlab.com/prvInSpace/ master-dissertation/-/blob/master/train/scripts/extract.awk

Two Python scripts were developed to split the data: one that simply splits the validated data[3] and one that splits it so that one sentence only appears in one of the splits[4]. The results from this test can be seen in table 5.1.

Model	Sentence		Utterance	
	WER	CER	WER	CER
EN-BR	80.71%	29.37%	83.15%	31.73%
EN-CY	65.11%	19.39%	65.69%	19.70%

Table 5.1: Summary of the models' performance depending on whether the datasets were split based on utterances or sentences.

Based on the results in table 5.1, it can be concluded that it does not make a significant difference whether the dataset is split based on utterances or sen- tences, at least for Welsh. The differences between results for the two approaches are minimal as to be within a margin of error. There is a marginal improvement for Breton, however, in the opposite direction of what was expected. This is probably due to the Breton dataset being small which makes it more suscepti- ble to "lucky splits". Since the worry was that the dataset containing the same sentences could give the models an unfair advantage, this result can simply be discarded.

Despite these results, the rest of the models will be trained using the data splits created by ensuring that each sentence is only present in one data split. The reason for this is to be on the safe side and to ensure that the models are not benefiting from testing on seen data. This is despite these results showing that there is no tangible benefit from training on unique utterances.

5.3 Training the models

This section describes the work that was undertaken to train the two base models and the six target models. It also describes the choices that went into selecting and optimising the hyperparameters.

5.3.1 Selection of an English base model

An English model for Coqui already exists and is released alongside every Coqui release. Since a workable English model already exists, it was decided that it was unnecessary to create another English model from scratch and that this model could be used instead. Since Coqui 1.3.0 was used and was the latest

[3]The Python script for splitting the dataset based on utterances can be found at https://gitlab.com/prvInSpace/master-dissertation/-/blob/master/train/scripts/ split_utterances.py

[4]The Python script for splitting the dataset based on sentences can be found at https://gitlab.com/prvInSpace/master-dissertation/-/blob/master/train/scripts/ split_sentences.py

release available at the time, the English model that was released for version 1.3.0 was used[5].

Testing the English model

No data was available for the character error rate of only the acoustic model without a language model. Since this data is required for some of the compar- isons and the analysis later, the model had to be manually tested. The problem is that because of the size of the English model, it was unfeasible to test it the same way that the other models were tested. However, by exporting the English model to a Tensorflow Lite file (tflite), it was possible to load the model successfully and test it using Coqui's evaluate_exported.py script. The only issue with this approach is that this script does not print a loss value, hence for the results this field will be empty. Given that the loss does not really tell us that much about the performance of the model, this does not particularly matter.

The test set used for testing the model was the predefined test set in the English Common Voice dataset. There are a couple of things that are worth noting. Firstly, the dataset is massive compared to its counterparts. Given the number of sentences already available in the test set, there was not really a reason to create custom splits for this dataset.

The dataset also had some clear instances of vandalism where certain record- ings contained computer-generated speech that said something different than the transition. It is uncertain to what extent this is a problem and how many of the test files were affected by this, but it likely negatively impacted theresults of the tests. That being said, it does not seem to be a big enough problem to invalidate the results, but it is definitely something to be awareof.

5.3.2 Training the French models

There are several French Coqui models available with good character error rates (CER). Examples of these are the model by Commonvoice-FR contributors (2022) that has a CER of approximately 15% and the model by Bermuth et al. (2021) that has a CER of approximately 9%. However, many of these use a combination of several large corpora. To ensure that most of the models were trained and tested using the Common Voice dataset, it was decided to train two French models from scratch: one from scratch and one using transfer learning using Coqui's English model.

The model that used the English base model as a base trained successfully without any issues. The same could not be said for the French standalone model. Several attempts were made in an effort to get it to train properly. The original attempt which can be seen in figure 5.1 was clearly not optimal. When training a model, it is expected that the loss will fall quickly in the beginning as the model is learning rapidly, then it slows down as the training slows down, and

[5]The English model can be accessed from the Github page for Coqui STT alongside the 1.3.0 release of Coqui at https://github.com/coqui-ai/STT/releases/tag/v1.3.0

then plateaus. However, while the loss curve does to some extent follow that expectation, it is flatter and in the end, it overfits completely.

Several methods and parameters were tested but none yielded the wanted results. In the end, it was decided to both lower the learning ratesubstantially to 0.0001 and increase the dropout rate to 0.3 which are the same parameters as used by Commonvoice-FR contributors (2022).

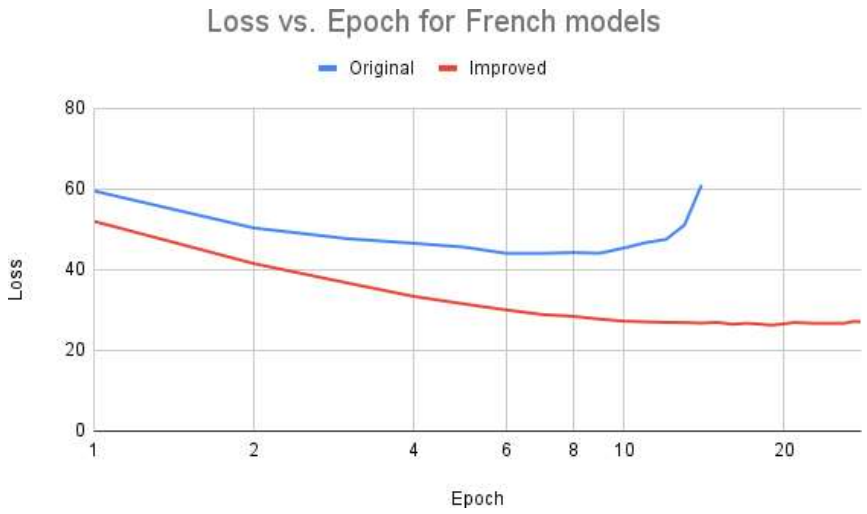

Figure 5.1: Loss vs. epoch for different French models

This improved model had a much better loss curve than the original at- tempts. This loss curve can also be seen in figure 5.1. Compared to earlier attempts, this model also had a significantly better performance. Full results will be discussed in section 5.4, but the French standalone model had a charac- ter error rate of 11.9% while the French model based on the English model had a character error rate of 14.9%. There are some interesting observations that can be made by looking at these results. Most significantly is that the French standalone model performed significantly better than the one made using trans- fer learning. This is something that has been observed in other domains and contexts before, notably by Virtanen et al. (2019).

5.3.3 Training models based on the English model

All of the models trained using the English (EN) model and English-French (EN- FR) model used Coqui's default settings. This means a dropout rate of 0.05 and a learning rate of 0.001. This seemed to yield effective results, comparable to other models for the same languages. Therefore, no further hyper-parameter optimisation was carried out.

5.3.4 Training models based on the French model

When models were trained using Coqui's default parameters, the models were unable to train properly. They followed an expected learning curve for a couple of epochs and then got rapidly worse. Different parameters were tested, but the ones that seemed to give the best results were a learning rate of 0.0001 and a dropout rate of 0.3. Hence, all models trained on the French base model used these parameters.

5.3.5 Overview of the loss during training

Following the optimisation of the hyper-parameters, all of the models trained mostly as expected. The loss curves for the training processes also follow the standard pattern to a satisfactory extent. The loss curves for all of the models can be seen in figure 5.2.

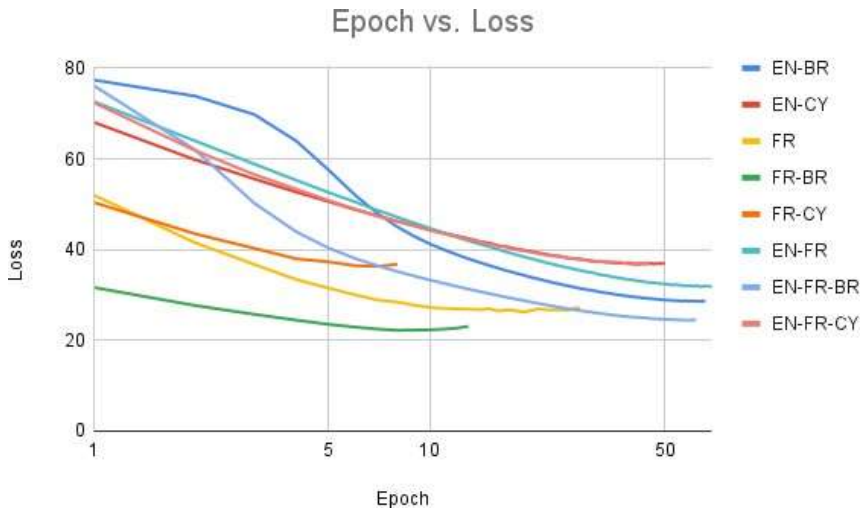

Figure 5.2: Overview of the loss curves for the models

One thing to note with the results shown in figure 5.2 is the considerable difference in training time between the models. Notably, the training times for models trained using the French base model were considerably shorterthan their English-based counterparts. Why the training times are so much shorter is unknown, but it is so frequent and consistent that it could warrant further investigation in the future.

5.4 Evaluation

All of the models were tested using the evaluation script that comes with Coqui along using the designated testing sets. The call to the evaluate script that was used can be seen in snippet 5.3.

Snippet 5.3: eval.bash

```
1 python3 -m coqui_stt_training.evaluate \
2         --test_files /data/clips/prv_test.csv \
3         --test_batch_size 64 \
4         --alphabet_config_path /data/alphabet.txt \
5         --checkpoint_dir /export/checkpoints \
6               | tee /export/results/\$(date + F-?R).txt
```

The parameters seen in snippet 5.3 follow the same naming convention as those used in the training script. The --test_files flag (line 2) is used to provide the files to use for testing. The --test_batch_size flag (line 3) is used to enable the script to process 64 files at a time instead of 1. The --alphabet_config_path flag (line 4) is used to specify what charac- ters the program is expected to find in the testing files. And finally the --checkpoint_dir(line 5) is used to specify which model to test.

The output of this script is then piped to the Unix program tee which is used to split the output of the evaluation to STDOUT and into a separate file so that it could be accessed later. The filename is based on the current time so that it will have a unique name.

5.4.1 Overview of the results

The word-error-rates (WERs), character-error-rates (CERs), and the loss value from the evaluations can be seen in table 5.2. The models are named according to which languages they are based on, so the models that are based on the English model start with “EN”, the multilingual model “EN-FR”, and so forth.

Model	WER	CER	Loss
EN	0.537	0.238	N/A
EN-BR	0.807	0.294	26.84
EN-CY	0.651	0.194	35.71
FR	0.369	0.119	26.45
FR-BR	0.655	0.222	20.24
FR-CY	0.604	0.189	35.60
EN-FR	0.467	0.149	31.87
EN-FR-BR	0.732	0.243	22.20
EN-FR-CY	0.640	0.194	35.88

Table 5.2: Overview of the word-error-rates (WER), character-error-rates (CER), and loss for the different models

Since no language models were used, the most interesting metric is the character-error-rate (CER) as this represents the models' ability to produce the correct sequences of letters the best.

As can be seen in table 5.2, the results varied quite a bit for the Breton model while the Welsh models stayed pretty consistent regardless of the base model. How these results compare to the state-of-the-art will be discussed in section 5.5.

5.4.2 Comparing the results against the hypothesis

At first glance, it is clear that the Breton models saw a significant improvement in the CERs and WERs when using the French models as opposed to the English model.

Contrary to expectations which would stipulate that multilingual models provide a better foundation for transfer learning, the results show that the monolingual French model outperformed the English-French multilingual one. Considering that this is only a single data point, it is not possible to draw any firm conclusions based on these results, but they do raise some interesting ques- tions. For example, given the performance of the monolingual model, it might be that multilingual transfer learning has an adverse effect on the performance of the model. It might also be that the performance of the base model plays a role and that this is something that should be explored further.

However, to show whether there is any merit to the hypothesis that the phonemic differences between of the base language and the target language can impact the quality of the model the CERs of the models need to be compared against the metrics that were extracted in section 4.3. This includes looking at whether there is a correlation between the CERs of the models and the phoneme overlap and Euclidean distance between the languages.

Comparing the CERs against the phoneme overlap

When we plot the CERs of the models against the percentage of phonemes in the target language present in the base language, we get the plot in figure 5.3.

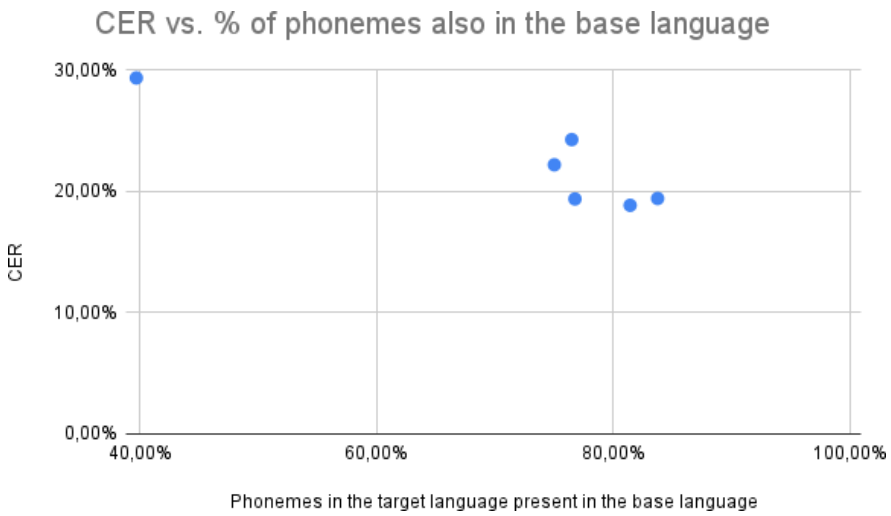

Figure 5.3: Plot showing the relationship between the CERs of the models and the percentage of phonemes in the target language present in the base language. Pearson's correlation coefficient for this relationship is r =–.908 which is a very strong negative correlation. The statistical significance of this correlation is p = 0.012 which means that it is statistically significant. This seems to underpin the hypothesis as if the languages have a higher overlap, we would expect a lower CER and this shows that.

That being said, given the extreme outlier that is the Breton model based on the English model, it might very well be that this makes this correlation much higher and much more statistically significant than reality. Pearson's correlation coefficient is known to not be robust when extreme outliers are present (Devlin et al., 1975). Given that the outlier is caused mainly by the stark difference between Breton and English, it does not make sense to remove it in this context. In order to properly investigate whether this data point is causing Person's correlation coefficient to overestimate the correlation, further data points would have to be added.

Comparing the CERs against the Euclidean distance

Another metric that was gathered in 4.3 was the Euclidean distance between the relative frequencies of the phonemes between the base language and the target

language. A plotted graph showing the correlation between the character error rate and the euclidean distance can be seen in figure 5.4.

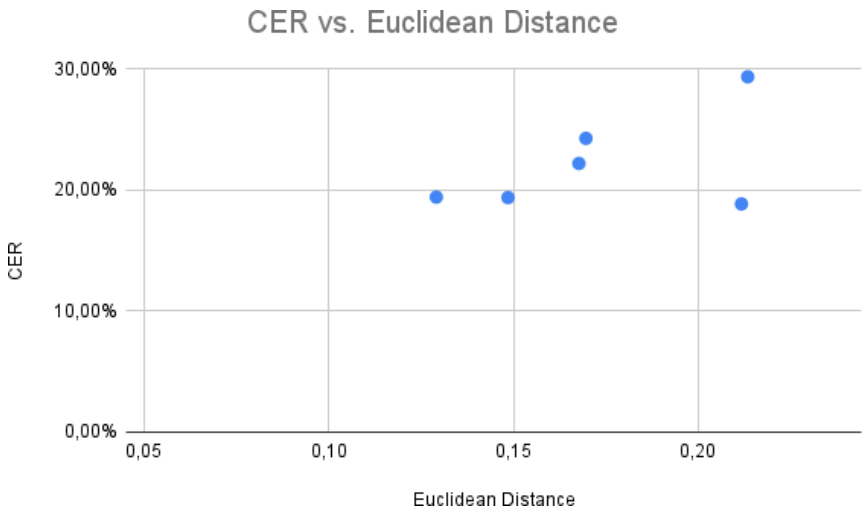

Figure 5.4: Plot showing the relationship between the CERs of the models and Euclidean distance between the relative frequencies of phonemes in the base language and target language

Pearson's correlation coefficient for this relationship is $r = .502$, however, the statistical significance is $p = .210$, meaning that this correlation is not statistically significant. As can be seen in 5.4, there does not seem to be a strong correlation.

Comparing the CERs against the number of unseen phonemes

The next relationship we will look at is the relationship between the CERs of the models compared to the number of unseen phonemes. This refers to the absolute number of phonemes in the target language that the base model have not seen before. This metric is quite similar to the percentage of phonemes in the target language that is also present in the base language, however, it does differ in some key aspects. The most significant difference is that this metric uses the absolute number of unseen phonemes as opposed to a percentage of phonemes present. The relationship between CERs and the number of unseen phonemes can be seen in figure 5.5.

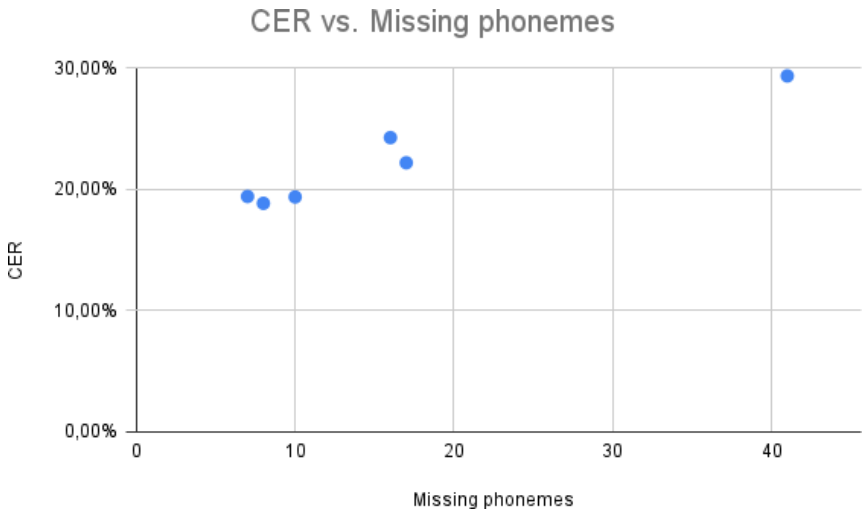

Figure 5.5: Plot showing the relationship between the CERs of the models and the number of unseen phonemes

of this relationship is very strong at $r = .961$. The statistical significance of this $p = .002$ which since it is lower than 0.05 makes this statistically significant. Using Pearson's correlation coefficient we can calculate that the correlation

That being said, similar to other metrics this suffers from a notable outlier which probably inflates the correlation and the statistical significance signifi- cantly. While it appears that there is a stronger likelihood of a correlation than previous metrics, most of the points are however between 20% to 25% CER and between 10 to 20 unseen phonemes. Given this cluster, it is possible that this correlation goes away if further data points are added.

Comparing the CERs of the models against the CERs of the base models

There is another way of analysing at these results which may explain these re- sults. We can look at the CERs of the base models and how these are correlated to the CERs of the Breton and Welsh models. If we plot this, we get thegraph that can be seen in figure 5.6. The data shown in figure 5.6 clearly shows that there is a correlation between the CERs of the base models and the CERs of the Breton models. While the Welsh model that is based on the French base model did perform marginally better than the others, the same trend can not be seen in the CERs of the Welsh models.

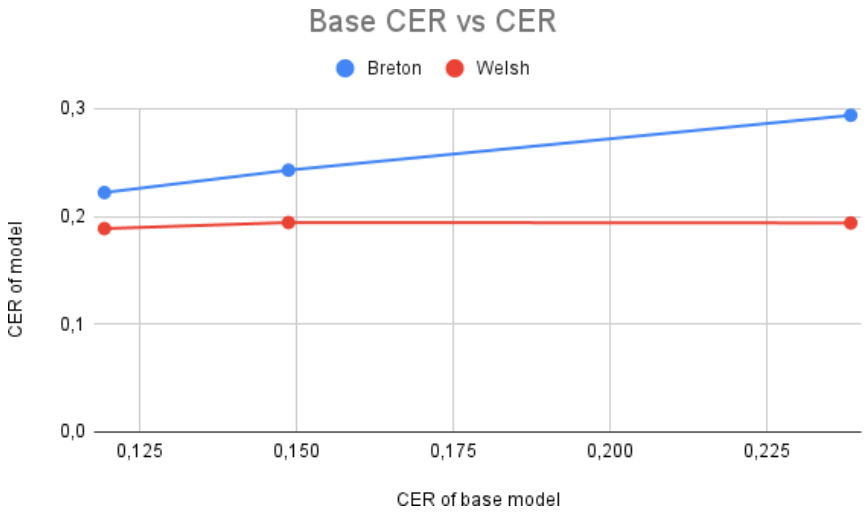

Figure 5.6: Graph showing the relationship between the CER of the base models vs. the CER of the Breton and Welsh models.

This raises some questions as to whether the original hypothesis is true or whether the underlying cause for the better performance is actually due to the better base model used. The issue is that there are not enough data points to draw any conclusive results. The only way of confirming whether this would be the case is to extend the scope of the project and look at other languages as well.

5.5 Evaluating the models up against existing models

To put these results into perspective, we will look at how they compare to other contemporary results and how they compare to the state-of-the-art for these languages.

5.5.1 Evaluating the Welsh models against existing models

The models in Jones (2022) were tested using the Bangor University Language Technology Unit's test set called the Corpws Profi Adnabod Lleferydd (Speech Recognition Test Corpus)[6] (Jones et al., 2022). Farhat (2022) also developed several acoustic models for Welsh and used the same testing set as Jones (2022).

[6]The dataset is available at the Language Technology Unit's own Gitlab in- stance at https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/ corpws-profi-adnabod-lleferydd/-/tree/master/

In order to beable to compare how the models developed for this project com- pares to those models, it was decided to test the models using this dataset as well.

The contents and the development of this dataset are explained further by Jones (2022). An important thing to note about this dataset is that it is non- verbatim. That means that grammar mistakes in speech have been corrected. This makes the corpus less ideal for the purposes of testing a standalone acoustic model, but despite this, it'll make comparing the effectiveness of the different models easier.

The Welsh model that was based on Coqui's English model managed to get a word error rate of 80.32% and a character error rate of 34.12%. This outperforms the model created by Jones (2022) which had a word error rate of 92.32% and a character error rate of 43.26%. It also outperforms the model created Farhat (2022) which had a word error rate of 82.82% and a character error rate of 38.52%.

It should be noted that both Jones (2020) and Farhat (2022) used older versions of Common Voice, and that more data has been released since they trained their models. Jones (2020) also developed several wav2vec2 (Baevski et al., 2020) models that outperformed both their Coqui models, but also the Welsh models trained for this project. It is clear, however, that transfer learning did significantly benefit the models and better results were achieved using this approach.

5.5.2 Evaluating the Breton models against existing mod- els

The biggest improvement can be seen in the Breton models. The best publicly available Breton models are the ones created by Tyers and Meyer (2021). Their Breton model had a CER of 41.56% which was later improved to 37.71% by optimising the hyperparameters. The three Breton models produced for this project all performed better than this, and the Breton model that was trained using the French monolingual base model performed the best achieving a CER of 22%. Some of this improvement can be explained by the fact that Tyers and Meyer (2021) used Common Voice 6.1 and this project used Common Voice 10 which includes new data. However, it is clear that transfer learning did benefit the Breton models, especially when a French base model was used.

5.6 Summary and discussion

Overall, the majority of the results from this experiment were inconclusive. While both the number of unseen phonemes and the percentage of phonemes present in the target language also present in the base language metrics showed a strong and statistically significant correlation when compared to the CERs of the models, it is clear that these can not be fully trusted. Given the Pearson's correlation coefficient's lack of robustness and the significant outliers, it is hard

to determine whether the trends that could be seen in the data were due to noise or a genuine correlation. More data is required to definitively determine whether there is any substance to the original hypothesis.

It is clear that transfer learning did benefit the models, and all of the models produced for the project were either comparable with state-of-the-art or substantially better. This is especially true for Breton, where the models produced for the project reduced the state-of-the-art character error rates from 37.71% to 22%. This improvement is important because it makes efficient speech-to-text and transcription services possible for Breton, a language which has historically been under-resourced.

Chapter 6

Training and Evaluating Further Models

Due to the results from the experiment yielding inconclusive results and correla- tions with questionable robustness, it was decided to expand the analysis further by running the experiment with some additional languages. This chapter will start off by reviewing what languages were selected and the reasons why. After that, the chapter will explain how the metrics extracted in 4 were extracted for Portuguese. Following that, the chapter will summarise the training process and describe any issues that were encountered. The chapter will evaluate these models and compare them against the previous results. By doing this, we can determine whether the results found in the last chapter are genuine correlations or simply a case of random chance. Once we have all of the results and the impact they have on the previous findings have been determined, the chapter will also go further in-depth to try to explain some of the findings that have been uncovered.

6.1 Review of languages selected

To improve the robustness of the results, it was important to increase the number of base languages as well as the number of target languages. Training a base model is a tedious process due to the amount of data available. Therefore it was decided that only adding one additional base language was the most feasible approach. This would also free up time to add several additional target languages as these have significantly shorter training times.

When it came to the selection of base language, it was decided to choose a language that was a bit further removed from Welsh and Breton than both French and English. However, the number of languages on Common Voice that had a suitable amount of data was low and as such it was mostly a choice between languages like Catalan and German. Since Catalan is fairly closely related to French, it was decided to use German as the base language for this expanded

experiment. German has about 1389 hours of validated data on Common Voice, which is significantly more than French but is feasible to train the models within a reasonable time.

When selecting the new set of test languages, it was important to keep two things in mind. Firstly, the language needed to be a lower-resourced language, and secondly, it needed to have a varied connection to the base languages. Three languages were chosen from the ones that were available on Common Voice.

The first one was Romansh. Romansh is a Romance language located in Switzerland meaning that it is related to languages such as French and Ital- ian. Despite this, it is a unique and distinct language with many influences from languages like German. Given that we have both French base models and German base models, this language is in a unique position to provide some interesting data due to being related to and influenced by both. Romansh is usually split into several different dialects, two of which are available on Com- mon Voice. These are Sursilvan Romansh which has approximately 6 hours of validated data while the other is Vallader Romansh which has about 2 hours. Because of its larger data size, it was decided to choose Sursilvan for this ex- periment. There is, however, an argument that both could have been used, but for simplicity's sake, it was decided to only use one. Another interesting idea for future research would also be to investigate transfer learning entirely within the Romansh context to see if different dialects could all benefit from the larger amount of data available for the Sursilvan dialect.

The second language chosen was Portuguese. Portuguese is interesting be- cause it is one of the few Romance languages except French to have nasalised vowels (Cruz-Ferreira, 1995). In addition to this, it also has a generally high phonemic overlap with French. In terms of hypothesis, this language is interest- ing, because if the hypothesis is true it is likely that we would expect Portuguese to benefit from using a French base model as opposed to both English and Ger- man. Portuguese has 120 hours of validated data on Common Voice making it the largest of the target languages for this experiment having narrowly more than Welsh.

The third and final language is Galician. Galician is a language from the Galicia region of Spain. It is very closely related to Portuguese and shares many linguistic and phonetic traits, and the two languages form a dialect continuum on the western side of Iberia. Compared to Portuguese, Galician is much more under-resourced having only 11 hours of validated data on Common Voice. This means that it has slightly more hours than Breton, while Portuguese has ap- proximately the same number of hours as Welsh. Given the similarities between these two languages, any difference in the performance of models trained with these two languages would likely be due to the additional data. This makes it interesting because it provides a greater view into how the amount of data impacts the models.

6.2 Extraction of metrics

One of the issues with Galician and Romansh is that due to the limited amount of resources available for these languages, there are not any pronunciation dic- tionaries available online. This means that it is extremely challenging to do the same analysis as was carried out in the last chapter using these twolanguages. This problem was known when selecting the languages and effectively means that they can not be used to evaluate the original hypothesis. That being said, they do provide valuable insight in other contexts such as when measuringthe impact of the base model's CER on the CER of the models.

Pronunciation dictionaries do exist for both Portuguese and German. The German pronunciation dictionary that was used was made by Minixhofer (2019) and contains about 365k words in IPA. This dictionary uses standard German for its transcriptions.

The Portuguese dictionary that was used was made by Mendonça and Aluí- sio (2014). There are some points to note about this dictionary, however. Por- tuguese, since it is spoken across the world in different countries, have many of distinct varieties. Most notably for our purposes are Brazilian Portuguese and Portuguese Portuguese. The pronunciation dictionary used for this project used the Brazilian variety and is centred around the dialect spoken in São Paulo. Since there is a difference in how Brazilian Portuguese and Portuguese Portuguese are pronounced, this means that we might lose some information. Hopefully, this should not impact the results too drastically.

Overview of the extract

Before we look at the results, let us look at how well the script was able to convert tokens to IPA and how many phonemes it found for each language. A summary of this can be seen in table 6.1.

Language	% Converted	Phonemes found
Breton	72.77%	68
English	96.36%	38
French	87.73%	82
German	66.21%	80
Portuguese	48.60%	37
Welsh	98.39%	43

Table 6.1: An updated overview of how many phonemes were found and how many tokens were successfully converted to IPA.

As can be seen in table 6.1, the script performed substantially worse at converting German and Portuguese to IPA in contrast to the previous languages. This is especially true for Portuguese where less than 50% of all tokens were successfully converted. Since all of the tokens were converted to lowercase so as

to make the conversion case-insensitive, this indicates that many of the tokens were not present in the pronunciation dictionaries or that the tokenisation was less than optimal. In the case of Portuguese, it is unlikely that poor tokenisation is the problem. Portuguese is written in a similar way to most other European languages using the Latin alphabet. Therefore there should not be any major difference between Portuguese and other languages. Hence, it is likely that a large number of tokens were unable to be converted properly. Whether this is because of missing diacritics or a difference in written standard is difficult to say without further investigation. The conversion rate for German is also noticeably lower than the other languages. Again, this is likely due to a failure to look up the tokens properly. To definitively determine whether this is because of a difference in written standard, missing tokens or something else requires further investigation. Despite these issues, however, there should still be enough data for both languages to base the analysis on.

One thing to note as well is the lower number of phonemes for Portuguese. Similar to English and Welsh, it is likely that Portuguese is more broadly tran- scribed than its counterparts, meaning that allophones are represented by one phoneme as opposed to multiple which would have been the case if the words were narrowly transcribed. This might have contributed to these results being more unreliable than what would otherwise have been the case.

Overview of extracted results

The extraction of metrics was carried out the same way as in chapter 4, and the same scripts were used. This produced the results that can be seen in table 6.2.

Target	Base	Phoneme overlap	Euclidean distance	Missing phonemes
br	de	79.41%	0.1721	14
cy	de	83.72%	0.1668	7
pt	de	91.89%	0.1893	3
br	en	39.71%	0.2134	41
cy	en	76.74%	0.1485	10
pt	en	64.86%	0.2320	13
de	en	46.25%	0.1563	43
fr	en	45.12%	0.2413	45
br	en-de	80.88%	0.1892	13
cy	en-de	86.05%	0.1366	6
pt	en-de	94.59%	0.2085	2
br	en-fr	76.47%	0.1696	16
cy	en-fr	83.72%	0.1291	7
pt	en-fr	89.19%	0.1867	4
br	fr	75.00%	0.1677	17
cy	fr	81.40%	0.2117	8
pt	fr	89.19%	0.1696	4

Table 6.2: Table showing all of the extracted metrics for all models created for the project

The results as seen in table 6.2 generally follow expectations, though there are a couple of things to note. German seemed to be a very good phoneme overlap with both Welsh and Portuguese. This was in many ways surprising, but likely boils downs to the fact that there were a lot of phonemes extracted from the German dataset meaning the chance of overlap is quite high.

If the hypothesis is true, then if the phoneme overlap and the absolute num- ber of missing phonemes are the most important metrics then German would be the best language to base Portuguese on. Otherwise, if the Euclidean distance is the most important, French would be the best. This was fairly unexpected.

For Welsh and Breton, the situation also changes. Both of these languages have a high phoneme overlap and a low number of unseen phonemes when compared to German and are likely to perform well using a German model if the hypothesis is true.

6.3 Training the models

The models were trained in the same way as described in section 5.3. Three models were created for each language, one based on Coqui's English model, one on the French model, and one on the French model which is based on the English model. This increased the total amount of models from 6 to 15 in order to provide some more data points.

6.3.1 Training the base models

The German model trained without any significant issues. By this point, how- ever, it was becoming increasingly clear that most models performed better when the learning rate had been reduced and the dropout rate had been increased. This unfortunately raises the possibility that some of the earlier models are not well optimised.

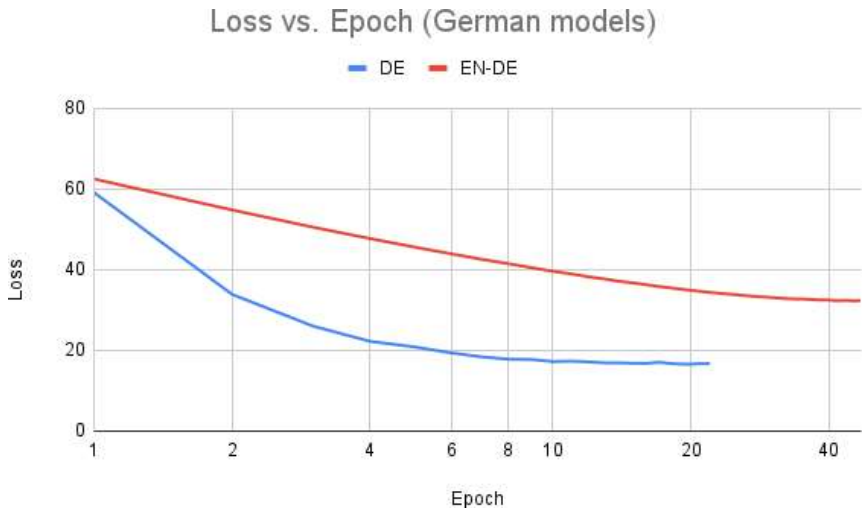

Figure 6.1: Overview of the loss curves for the two German base models

The loss curves for the two German base models can be seen in figure 6.1. As can be seen on that graph, the English-German model took significantly longer to converge than its monolingual counterpart. (Please note the log scale on the x-axis).

6.3.2 Training the other models

The training of the Breton, Welsh, Romansh, and Galician models had no major issues, and they all resulted in adequately efficient models. They were all trained using a dropout rate of 0.3 and a learning rate of 0.001.

Portuguese, however, was more difficult to get right. Firstly, due to issues with the dataset, there are entries that do not seem to work properly. This means that it is not possible to split the validated set into custom splits using the original scripts that were developed. Hence, the predefined splits, which might be less than ideal since they are missing a lot of data, had to be used.

The second issue is due to the poor performance of the models. Regardless of the parameters used, the models did not improve much further than the first

attempts. Only lowering the learning rate to 0.0001 made a difference, but it also increased the training time significantly for almost no benefit. A quick look at the data does provide some potential reasons for the poor performance of Portuguese. Based on the accent tags in the dataset, it seems as if the training set is exclusively made up of Brazilian Portuguese, while only the testing set and to a certain degree the validation set has any Portuguese Portuguese data. Due to the majority of entries not being tagged, it is hard to know for certain whether this is the case, but the tags that are there seem to suggest that the sets are quite imbalanced.

6.4 Evaluation with the new models

This section will look at the same relationships as those that were looked at in the previous chapter, but with the new and extended dataset. The goal is to uncover whether the results in the previous chapter actually hold up to scrutiny or whether any perceived correlation is simply due to random chance.

6.4.1 Overview of results

Including the six original target models, a total of 15 target models were trained. These can be seen in table 6.3 alongside the 5 base models. When analysing such a large table, it is difficult to ingest what the most notable entries are. Therefore we will briefly analyse how these results impact the results from the original analysis, how they compare to the state-of-the-art and other aspects that are of interest.

Model	WER	CER	Loss
DE	29.12%	7.01%	16.47
DE-BR	65.75%	22.51%	20.84
DE-CY	58.88%	18.19%	34.07
DE-GL	78.71%	23.14%	38.31
DE-PT	72.29%	26.23%	36.40
DE-RM	72.77%	19.78%	36.42
EN	53.68%	23.83%	N/A
EN-BR	80.71%	29.37%	26.84
EN-CY	65.11%	19.39%	35.71
EN-DE	53.32%	16.00%	35.81
EN-DE-BR	70.24%	23.67%	20.80
EN-DE-CY	54.68%	16.01%	28.87
EN-DE-GL	76.25%	21.02%	34.51
EN-DE-PT	71.94%	24.92%	33.74
EN-DE-RM	68.98%	17.51%	31.19
EN-FR	46.70%	14.87%	31.87
EN-FR-BR	73.16%	24.28%	22.20
EN-FR-CY	64.04%	19.43%	35.88
EN-FR-GL	85.98%	26.19%	45.49
EN-FR-PT	83.63%	31.53%	44.33
EN-FR-RM	83.67%	26.84%	49.63
EN-GL	91.23%	30.44%	53.31
EN-PT	83.80%	31.28%	43.29
EN-RM	91.84%	32.41%	57.62
FR	36.86%	11.93%	26.45
FR-BR	65.47%	22.21%	20.24
FR-CY	60.44%	18.86%	35.60
FR-GL	84.35%	26.24%	46.65
FR-PT	73.87%	26.69%	37.82
FR-RM	88.05%	27.94%	50.41

Table 6.3: A summary of the performance of all of the models trained for the project

6.4.2 Comparing the CERs against the phoneme overlap

significance below 0.2. The extended data can be seen in Figure 6.2. The first analysis this showed a strong negative correlation at $r_s = -0.906$ and a statistical metric that we will analyse is the phoneme overlap. In the original

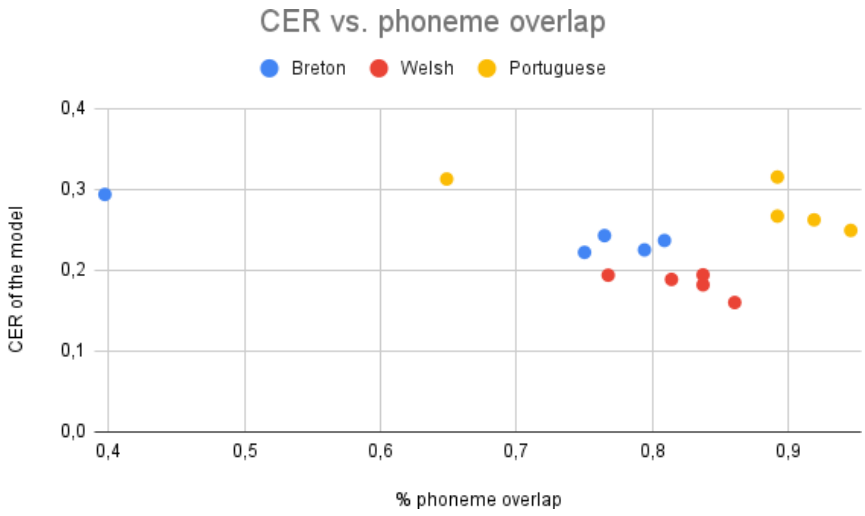

Figure 6.2: A plot showing the relationship between the CER of the models and the phoneme overlap.

As can be seen in figure 6.2, it is more clear now than in the original set of data that there is not a strong correlation between the CER of the models and the phoneme overlap between the target language and the base language. Calculating the Pearson's correlation coefficient for this relationship, we find

that the correlation is r = .159 and a statistical significant of p = .541. This effectively shows that the original results were inflated due to the non-robustness of Pearson's correlation coefficient. Not only that, it is undoubtedly a null result and it disproves the original hypothesis.

6.4.3 Comparing the CERs against the Euclidean distance

The second metric that we will analyse is the Euclidean distance between the base language and the target language. when the original analysis was carried out, this relationship showed a clear null result, having a correlation coefficient

of r = .502 and a statistical significance of p = .310. A plot showing the new data can be seen in figure 6.3.

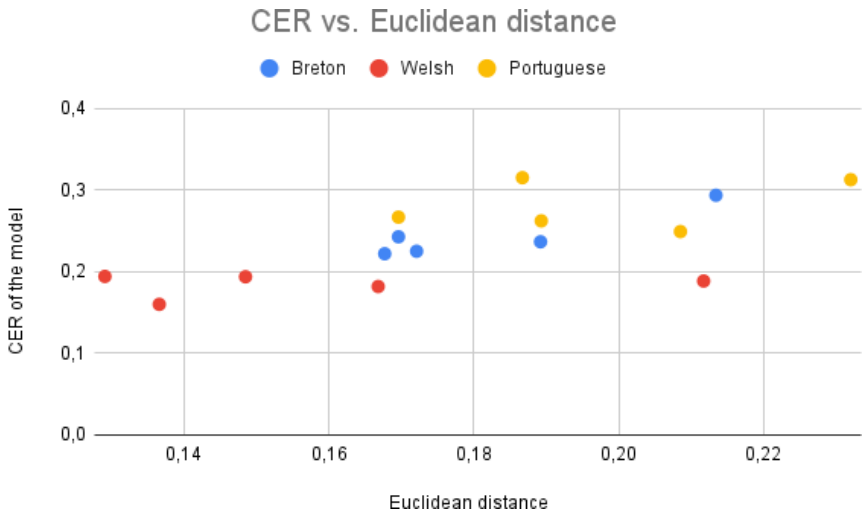

Figure 6.3: A plot showing the relationship between the CER of the models and the Euclidean distance.

Examining the results as seen in 6.3, it is evident that the original results still stand and that there still is not a statistically significant correlation. While it appears that there might be an upwards trend, when looking at individual lan- guages, it can plainly be seen that the results are to a certain extent random in nature. When calculating the Pearson's correlation coefficient for this relation- ship, we find that it has a correlation of $r = .360$ and a statistical significance of $p = .156$. This correlation is very weak and not statistically significant and shows that the original results do not hold up to more in-depth scrutiny.

overlap discussed above, this disproves the original hypothesis

Despite that, it should be noted that these results were the closest to showing a correlation. However, it is unlikely that additional data would change the outcome of this, but it is worth keeping in mind for further investigation in the future.

6.4.4 Comparing the CERs against the number ofunseen phonemes

The last of the main metrics that we will examine is the absolute number of unseen phonemes (i.e the absolute number of phonemes present in the target language that is not present in the base language). The original analysis showed a Pearson's correlation coefficient of $r = .961$ and a statistical significance of $p = .002$. While the original plot was the most promising of the original set of relationships, it was transparent that there were significant outliers that were likely inflating both

correlation was the strongest correlation having a Pearson's correlation coefficient of $r = .961$ and a statistical significance of $p = .002$. While the original plot was the most promising of the original set of relationships, it was transparent that there were significant outliers that were likely inflating both

the correlation coefficient and the statistical significance of the relations. The updated plot with the new data can be seen in 6.4.

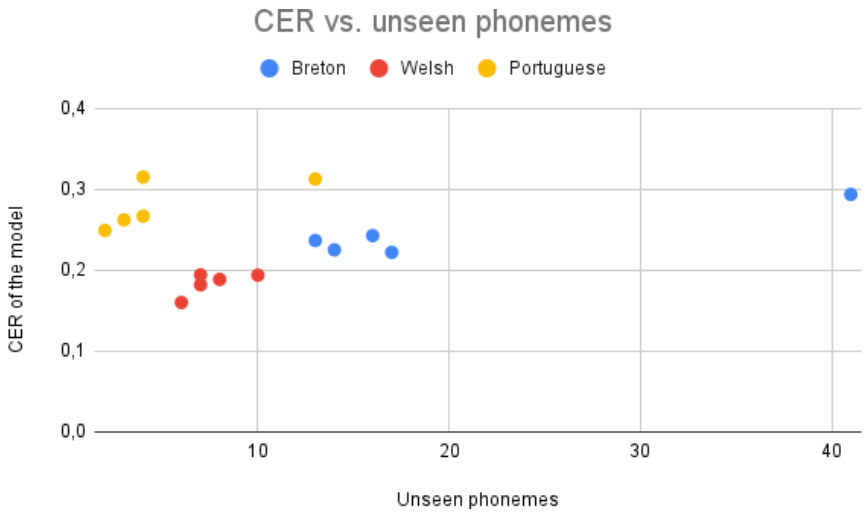

Figure 6.4: A plot showing the relationship between the CER of the models and the absolute number of unseen phonemes.

Even with the new data, there are still significant outliers as seen in figure 6.4. Despite the outlier, however, it still is not enough to skew the results to such an extent that it shows a correlation between the metric and the CER. Pearson's correlation coefficient for this relationship is $r = -.264$ and the statistical significance of this correlation is $p = .306$. This means that there is no statistically significant correlation and unexpectedly the original results did not hold up against further scrutiny. Similar to the previous two metrics discussed, this unquestionably disproves the original hy- pothesis and shows that there is no correlation between a model's performance and the base language used.

6.4.5 Summary of findings, limitations, and caveats

It is clear that all of the metrics result in null results and effectively disprove the original hypothesis. There are some caveats, however. Due to the nature of machine learning and the need for hyper-parameter optimisation, and random chance, it might very well be that some of the models are not as optimal as they could have been. Looking only at the result that Tyers and Meyer (2021) obtained, just optimising the hyper-parameters alone could improve the per- formance of models by up to 15%. While hyper-parameters have been tuned during this project in an attempt to yield the best possible models, there is

undoubtedly room for improvement.

There is also the issue of accuracy when it comes to the metrics chosen and how they were extracted. There is definitely a margin of error in these results and this might be quite substantial. So while all of the metrics resulted in null results, this should, in some ways, be taken with a pinch of salt. There are so many variables at play when dealing with machine learning, and results like the ones observed for Breton are hard to explain unless there is some characteristic with French that makes it such a good language to base Breton models on.

This should not be interpreted as there being any merit to the hypothesis, as based on the findings there is not. Nevertheless, it is important to stress that the results have an abnormally high margin of error.

6.4.6 Comparing the CERs against the CERs of the base models

One relationship that was of interest during the original experiment was the relationship between the CER of the base model vs. the CER of the target models. While this did not seem to affect the Welsh models, there was a clear linear relationship between the performance of the Breton models and the per- formance of the base models. The only way to investigate whether there is a genuine relationship is to gather additional data. By adding two German base models and three additional target languages, the number of data points has increased from six to 25. An updated plot showing the relationship with the new data can be seen in figure 6.5.

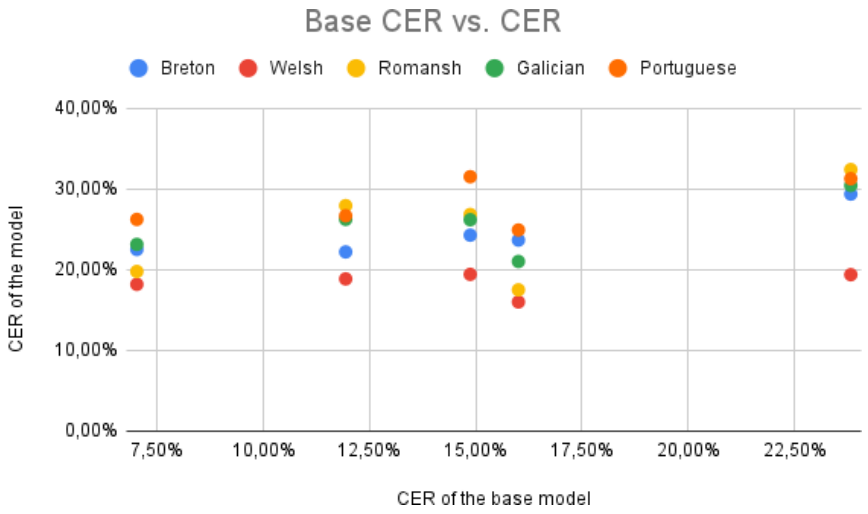

Figure 6.5: A plot showing the relationship between the CER of the base model and the CER of the target models

The relationship that can be seen in figure 6.5, contains some noteworthy features. Firstly, all languages, with the notable exception of Welsh, seem to follow the same pattern as Breton to a certain extent. It is also clear that the English-German model (second column from the right) sticks out promi- nently. Why this is the case is unclear. It might genuinely be that this model provided a well-balanced base model for most languages to train a model on. It might be that the training and testing sets accidentally got re-split before these models were trained. However, models based on the German model and English-German model were trained by language in pairs. Hence, if this was the case we would have expected the German results (furthest to the left) to also perform better, which they do not.

Pearson's correlation coefficient for this relationship is $r = .142$, which is a negligible correlation and the significance of this is $p = .479$, meaning that the correlation is not statistically significant.

There is one issue with this plot, however. Since the amount of training data varies substantially between the languages, comparing the raw CERs of the target models does not adequately compensate for this difference. In order to compensate for this difference, it was decided to normalise the CERs of the target models. This was done by setting the CER for the models trained using an English model to 1, then calculating the ratio compared to this for the rest of the models. That produced the plot that can be seen in figure 6.6.

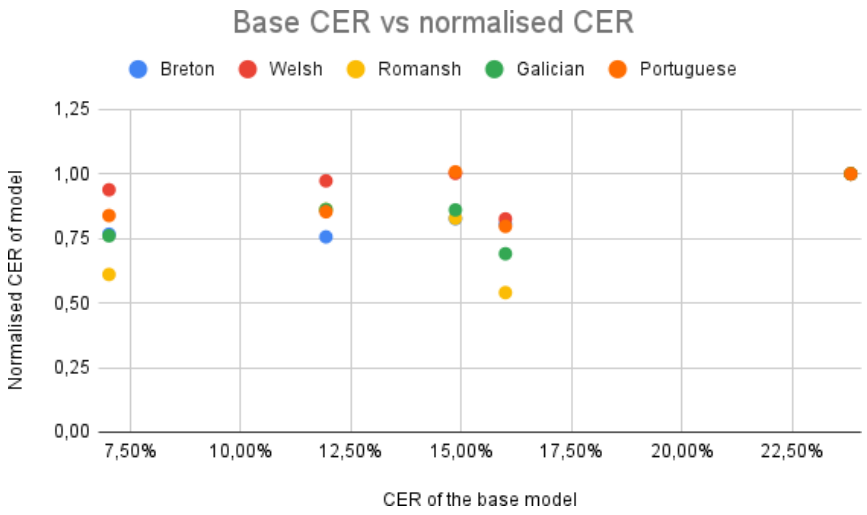

Figure 6.6: A plot showing the relationship between the CER of the base model and the CER of the target models

This normalised version of the plot reveals an interesting trend. Theredoes seem to be a significant positive correlation between the CER of the base models

relationship we find that the correlation is $r = .554$ which is a strong correlation. Not only that, the result is also statistically significant as $p = 0.0$ meaning that the correlation is statistically significant.

This is intriguing. There are a couple of explanations as to why this is the case. The most obvious one is that there are some intrinsic characteristics of the patterns that the model learns that are universally applicable. This makes sense in some ways because it is this universally applicable aspect that transfer learning is exploiting. Effectively, this would indicate that when a model becomes better at transcribing audio in the language it was designed for, it also becomes better suited and better capable of transcribing any audio to a certain extent.

The models trained using the English-German model still sticks out promi- nently. As explained earlier, there are many reasons why this might be the case, but there is no clear answer with the data that we have available.

6.5 How the models compare to contemporary models

This section will briefly go through models created for the new languages and see how they compare to the state-of-the-art for these languages.

6.5.1 German models compared to the state-of-the-art

There are two German Coqui models available from Coqui's websites, one by Agarwal and Zesch (2019) and one by Bermuth et al. (2021). Theperformance metrics for the model made by Agarwal and Zesch (2019) only lists the word error rate. Since this project has exclusively focused on the performance of the acoustic models and since the models do not have a language model, the results for this project is not comparable with Agarwal and Zesch (2019).

The German model trained for this project had a character error rate of 7.0%. The model trained by Bermuth et al. (2021) had a character error rate of 5.6% when testing on the Common Voice testing set. That means that the model trained for this project is slightly worse than the one trained by Bermuth et al. (2021). That being said, their model was trained using data from 17 different corpora in addition to the Common Voice dataset, so this is not very surprising.

6.5.2 Galician models compared to the state-of-the-art

There do not seem to be any available monolingual Galician models online. Both the models trained by Dieguez-Tirado et al. (2005) and Docio-Fernandez and Garcia-Mateo (2018) are Spanish-Galician bilingual models. As with some other models, they only list the word error rates which is not comparable in our

case. Therefore, there do not seem to be Galician monolingual models that are comparable to the ones developed here.

6.5.3 Portuguese models compared to the state-of-the-art

Portuguese was one of the languages that Tyers and Meyer (2021) used in their experiment. The best character error rate they got without the addition of a language model was 26.69% (They managed to get it down to 20.10% using a language model) The best Portuguese model trained for this project (EN-DE- PT) had a character error rate of 24.92%. This is comparable to the models trained by Tyers and Meyer (2021) and does not differ in any statistically sig- nificant way.

6.5.4 Romansh models compared to the state-of-the-art

Romansh was also one of the languages that Tyers and Meyer (2021) used in their experiment. In fact, they trained models for both the Sursilvan and Vallader dialects of Romansh. Due to Sursilvan having more data on Common Voice, it was decided for this project to focus on this dialect. The best model that Tyers and Meyer (2021) trained for the Sursilvan dialect without a language model had a character error rate of 23.88% (18.93% with a LM). The best Romansh model trained for this project (EN-DE-RM) had a character error rate of 17.51% which is not only significantly better than the model trained by Tyers and Meyer (2021), but also slightly better than their model with a language model. It is clear that Romansh did significantly benefit from transfer learning, and that effective models for the language are very much a possibility.

6.6 On the lasting impact of the base language on the models

Both the French and German models had a worse performance when transfer learning was used. The exact reason is difficult to definitively answer, however, it does seem to indicate that for non-lower-resourced languages, transfer learning can have an adverse effect on the performance of the models.

This is likely due to the presence of the base model, even after hours of training on a different large data set. Both German and French were trained on hundreds of hours of data, and German even over a thousand hours. Despite this, the training was unable to completely overwhelm the influence of the base model. This raises some interesting questions, like how much data is required for models to become adversely affected by transfer learning and whether there is a way to overcome this adverse effect. These are definitely questions that could be investigated by future research and could be beneficial for the field as a whole.

6.7 Summary and discussion

The experiment's scope was greatly expanded to improve the robustness of the results and limit the negative impact of the wide margin of error. By doing this, it was revealed that the original results did not stand up to additional scrutiny, and all of the statistically significant correlations found in the original analysis became not statistically significant. This underlines the importance of questioning the initial results and especially the robustness of the correlation calculations.

It was found that there is a statistically significant correlation between the base model's performance in its own domain and the performance of the mod- els that use this model as a base. This might be helpful to explain one of the contributing factors determining the effectiveness of transfer learning in a speech-to-text context.

Chapter 7

Conclusion and Future Work

In this chapter, we will review the work that has been undertaken and the find- ings of the dissertation. We will also look at the research questions, hypotheses, aims, and objectives to determine whether the dissertation has fulfilled its aims. The limitations of the work will also be laid out and future work will be dis- cussed.

7.1 Summary and conclusions

The dissertation first set out to quantify the relationship between different languages. This was achieved by extracting information about the phonemes present in the training data and comparing the results with other languages. When this work was carried out, it was also uncovered that the pre-defined data split in Common Voice makes poor use of the available data and that custom splits had to be created.

Several base models were trained for English and French, and these models were then used as a base for Welsh and Breton models to be trained on. Com- mon Voice was used as the data source for these models and Coqui STT was used has the STT framework. When the results were analysed, the dissertation orig- inally found some statistically significant correlations between the relationship between the languages and the performance of the models. However, when the experiment was expanded to include German, Romansh, Portuguese, and Gali- cian, these correlations went away. The dissertation has therefore been unable to show whether there is a correlation between the performance of the mod- els and the relationship between the languages. The dissertation has, however, shown that there is a statistically significant correlation between the character error rate of the base model and the character error rate of the complete models. It is clear that transfer learning substantially improves models for lower- resourced languages. The method used in this dissertation enables effective

models to be created for these languages without the vast amount of data that is available for other languages. This lowers the barrier of entry for tools and ser- vices to be developed for lower-resourced languages. Breton and Romansh saw significant improvement when compared to state-of-the-art while Portuguese saw a small improvement.

7.2 Review of research questions and hypotheses

There was one main research question that this dissertation set out to answer, and that was whether the relationship between the base language and the target language affects the performance of the complete models. This dissertation has shown that there is no statistically significant correlation between the perfor- mance of the models and the relationship between the languages. Therefore, the hypothesis (H_1) can not definitely be shown to be true. Additionally, this means that the null hypothesis (H_0), which states that the relationship between the languages does not impact the performance of the models, holds true based on these results. Given the abnormally high margin of error in this experiment, this does not definitively disprove the original hypothesis H_1, but the dissertation is unable to sufficiently determine whether there is a basis for the hypothesis.

Despite this, the dissertation has shown that there is a correlation between the performance of the base model and the target model. This shows that models that perform well in their own domain are better suited as base models to be used in transfer learning.

7.3 Review of aim and objectives

The aim of the dissertation was to improve speech-to-text models for lower-resourced language and to explore ways of optimising data utilisation for trans- fer learning. Looking at the performance of the models trained for this project, it is clear that transfer learning does in certain contexts substantially improve model performance compared to the state-of-the-art. For example, for Breton, the character error rates were lowered from 37% to 22% by using this method. When character error rates get down to that level, speech-to-text models be- come much more useful and practical. This not only significantly lowers the barrier of entry for technologies like speech-to-text for these languages, but they provide invaluable tools and services to different communities. Improving accessibility services for speakers of these languages is of vital importance and any improvement made to these services is of incredible value. It is fair to say that for especially Breton and Romansh, this dissertation succeeded in showing that transfer learning can be efficiently used to provide useful models for lower- resourced languages. In this way, the dissertation clearly fulfilled in achieving its aim.

There were also four concrete objectives that the dissertation set out to achieve. The first one was to create a way of extracting information about the

relationship between two languages. As described in chapter 4 this objective was successfully achieved.

The second one was to train bespoke and novel models for several lower-resourced languages using transfer learning. In total, 25 models were trained using transfer learning for several lower-resourced languages. For Breton and Romansh, the models trained significantly outperformed the state-of-the-art. For Galician, no available monolingual speech-to-text models exist, hence the models trained for this project provide a benchmark for any future development within this area. And for Welsh and Portuguese, the models were comparable with existing state-of-the-art. All in all, the project improved upon the state- of-the-art and created novel models for lower-resourced languages that did not have it before. Therefore, the project did succeed in fulfilling this objective.

The third objective was to investigate whether there is a correlation between the relationship between the languages and the performance of the models. Even though the project was unable to definitively determine whether there is a corre- lation or not due to the null results, the project did achieve this objective. More knowledge about what makes effective transfer learning models were uncovered. That in itself is an achievement in terms of the fulfilment of this objective, es- pecially since the knowledge gained from fulfilling objective three provided the basis for the fulfilment of objective four.

The fourth objective was to explore whether there are any other contributing factors that affect the performance of the models. The analysis carried out in this project uncovered that there is a strong correlation between the performance of the base model and the performance of the target model. This is interesting because it seems to suggest that a model's ability to perform its learning task T_s in its own domain D_s makes it better suited to be used as a base as it improves the target model's ability to carry out its learning task T_t in its domain D_t. There is definitively more work to be carried out with respect to objective four, but the work undertaken here provides a good basis for more work to be done in the future.

7.4 Limitations

This dissertation has some significant limitations. Firstly, only three base lan- guages and three target languages were tested. While this provided many data points, it is likely this is not enough to definitively answer the original research question due to the significant margin of error.

The second, and most important limitation, is the margin of error in these results. Due to the time-consuming nature of training and optimising models and the random element to training the models, it is likely that the models while good are sub-optimal. This makes any definitive analysis difficult, especially in the absence of more data, which of course would be extremely time-consuming to acquire. Add to that the margin of error introduced by the difference in how broad or narrow the transcriptions are, and it becomes even more difficult.

This is definitely the biggest limitation of the dissertation. Despite that,

the dissertation did go beyond the original scope in an attempt to lessen the effect of this, and by doing so made the results more robust and reliable. With more time and resources, fully optimising the models, improving the metric extraction process, and getting more reliable results could have been possible, but unfortunately the project had neither.

7.5 Future work

The results in the dissertation did in many ways not adequately and definitively answer the original research question. There is clearly a relationship between French and Breton that makes French better suited to be used as a base model for Breton than English. While the overall results are inconclusive and no statistically significant correlations could be shown, this trend was completely in-line with the original hypothesis and the original idea for the dissertation. There is definitely more work to be undertaken in this field, especially in relation to lower-resourced languages.

This dissertation focused entirely on the performance of the acoustic mod- els. Despite this, there are several questions that still remain in relation to how transfer learning can be used most effectively to aid in the creation of ef- fective models when language models are included. One of these questions is how to best utilise data available for one dialect of a language to aim in the development of models for a lesser-resourced dialect of the same language. It remains to be seen whether it is best to create bespoke models for each dialect or whether it is beneficial to share either an acoustic model, language model, or both. It might very well be that by using a combined acoustic model and a bespoke language model for each dialect you are able to more efficiently utilise the available data for these languages. This is something that existing literature does not adequately answer, and it is something that should be investigated in future research.

The field is moving forward, and effective models are becoming available for lower-resourced languages where those models were once considered impractical due to the lack of data. There is always room for improvement and room for new methodologies to be devised so that the limited data can be more effectively utilised. Therefore, the amount of future work required is vast and existing literature and this dissertation are only scratching the surface of the possible.

Bibliography

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

Aashish Agarwal and Torsten Zesch. German end-to-end speech recognition based on deepspeech. In *KONVENS*, 2019.

Alfred V. Aho, Brian W. Kernighan, and Peter J. Weinberger. Awk—a pattern scanning and processing language. *Software: Practice and Experience*, 9(4): 267–279, 1979.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gre- gor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representa- tions. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low- resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*, 2018.

JA Barnett. A phonological rules system. Technical report, System Development Corp, Santa Monica, CA, 1975.

Daniel Bermuth, Alexander Poeppel, and Wolfgang Reif. Scribosermo: Fast speech-to-text models for german and other languages. *arXiv preprint arXiv:2110.07982*, 2021.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Au- tomatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014.

Commonvoice-FR contributors. Common Voice Fr STT Model. https://github.com/wasertech/commonvoice-fr/releases/tag/v0.9.0-fr-0.1, 2022. [Online; accessed 27-September-2022].

Sarah Cooper, Dewi B. Jones, and Delyth Prys. Crowdsourcing the paldaruo speech corpus of welsh for speech technology. *Information*, 10(8), 2019. ISSN 2078-2489. doi: 10.3390/info10080247.

Coqui AI. Coqui stt, 2022. URL https://coqui.ai. Version 1.3.0.

Madalena Cruz-Ferreira. European Portuguese. *Journal of the International Phonetic Association*, 25(2):90–94, 1995. doi: 10.1017/S0025100300005223.

Susan J Devlin, Ramanathan Gnanadesikan, and Jon R Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62 (3):531–545, 1975.

Javier Dieguez-Tirado, Carmen Garcia-Mateo, Laura Docio-Fernandez, and An- tonio Cardenal-Lopez. Adaptation strategies for the acoustic and language models in bilingual speech transcription. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–833. IEEE, 2005.

Laura Docio-Fernandez and Carmen Garcia-Mateo. The gtm-uvigo system for albayzin 2018 speech-to-text evaluation. *Proceedings of the IberSPEECH, Barcelona, Spain*, pages 21–23, 2018.

Leena S. Farhat. Applying a teacher-student learning approach to improve Welsh STT frameworks. Master's thesis, Bangor University, August 2022.

Stephen J. Hannahs. French phonology and l2 acquisition. *French applied linguistics*, 16:50–74, 2007.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gom- mers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Se- bastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Shep- pard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585: 357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Pro- ceedings of the sixth workshop on statistical machine translation*, pages 187– 197, 2011.

Roparz Hemon and Michael Everson. *Breton grammar*. Evertype, 2011.

International Organization for Standardization. ISO 639-1:2002: Codes for the representation of names of languages—part 1: Alpha-2 code, 2002.

International Phonetic Association, International Phonetic Association Staff, et al. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

Dewi B. Jones. Macsen: A voice assistant for speakers of a lesser resourced language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 194–201, 2020.

Dewi B. Jones. Development and evaluation of speech recognition for the welsh language. In *LREC 2022 Workshop Language Resources and Evaluation Con- ference 20-25 June 2022*, page 52, 2022.

Dewi B. Jones and Sarah Cooper. techiaith/geiriadur-ynganu-bangor: Geiriadur Ynganu Bangor Pronunciation Dictionary, July 2021. URL https://doi. org/10.5281/zenodo.5112035.

Dewi B. Jones, Tegwen Bruce-Deans, and Stephen Russell. Corpws Profi Adnabod Lleferydd. https://git.techiaith.bangor.ac.uk/ data-porth-technolegau-iaith/corpws-profi-adnabod-lleferydd, 2022. [Online; accessed 27-September-2022].

Holly J. Kennard. Variation in Breton word stress: new speakers and the influ- ence of French. *Phonology*, 38(3):363–399, 2021.

Vladimir I. Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707– 710. Soviet Union, 1966.

Gustavo Mendonça and Sandra M Aluísio. Using a hybrid approach to build a pronunciation dictionary for brazilian portuguese. In *INTERSPEECH*, pages 1278–1282, 2014.

Dirk Merkel. Docker: lightweight Linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.

Christoph Minixhofer. German IPA Pronunciation Dic- tionary. https://www.kaggle.com/datasets/cdminix/ german-ipa-pronunciation-dictionary, 2019. [Online; accessed 27- September-2022].

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans- actions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glem- bek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. *arXiv preprint arXiv:1902.00193*, 2019.

Sebastian Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

Ilnar Salimzianov. A baseline model for computationally inexpensive speech recognition for kazakh using the coqui stt framework. *arXiv preprint arXiv:2107.10637*, 2021.

Patrick Sims-Williams. The Celtic languages. *The Indo-European Languages*, page 345, 2015.

John Tabak. *Geometry: the language of space and form*. Infobase Publishing, 2014.

Francis M. Tyers and Josh Meyer. What shall we do with an hour of data? speech recognition for the un-and under-served languages of common voice. *arXiv preprint arXiv:2105.04674*, 2021.

Preben Vangberg, Pierre Morvan, and Aodren Le Gloanec. Breton pronouncing dictionary, 2022. URL https://gitlab.com/prvInSpace/ breton-pronunciation-dictionary. [Online; accessed 27-September-2022].

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: BERT for finnish. *CoRR*, abs/1912.07076, 2019. URL http://arxiv.org/ abs/1912.07076.

Robert L. Weide, Alex Rudnicky, and Speech Group in the School of Com- puter Science at Carnegie Mellon University. Carnegie Mellon Pronouncing Dictionary. https://github.com/Alexir/CMUdict, 2015. [Online; accessed 27-September-2022].

Wikeriadur contributors. Wikeriadur, the free dictionary, 2022. URL https: //br.wiktionary.org/. [Online; accessed 27-June-2022].

David Willis. Old and middle welsh. *The Celtic Languages*, pages 117–160, 2009.

Mehmet Yavas. *Applied English Phonology*. John Wiley & Sons, 2020.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3(175):12, 2002.