Doing More with Less: Determining Optimal Pre-training Model for Irish Automatic Speech Recognition through Multi-step Fine-tuning

Caoilfhionn Ní Dheoráin, Ruth Holmes, Nicholas Evans, Thomas Laurent, Anthony Ventresque, Ellen Rushe
School of Computer Science and Statistics - Trinity College Dublin, SFI Lero,
School of Computing - Dublin City University
{nidheorc, ruth.holmes, nevans, tlaurent, anthony.ventresque}@tcd.ie, ellen.rushe@dcu.ie

Abstract

In recent years, there has been an upsurge in research on automatic speech recognition (ASR) for low-resource languages. Particularly, transfer learning using multi-lingual models has become a popular remedy for the lack of available datasets for target languages. However, given the complexities associated with each individual language, we argue it is unlikely that a single multi-lingual pre-training model will provide equal performance gains across all languages. We also recognise the important, and insufficiently studied influence that the specific pre-training dataset has on the performance of the model. In this paper, using the Irish language as a case study, we propose a more directed, incremental form of pre-training which we term *multi-step fine-tuning*. This method accounts for the complex relationships between the language and dataset features of the source pre-training and target datasets. We show multi-step fine-tuning improves performance over simple multi-lingual fine-tuning alone, and we investigate factors leading to certain pre-trained models achieving better results through linguistic and dataset similarity measures. This research also investigates the uniformity of the performance gains across different demographics. We show that the optimal pre-training strategy can differ between demographics suggesting that more careful pre-training dataset selection is necessary to ensure equitable outcomes in practice.

Keywords: Automatic speech recognition, low-resource language, language similarity, Irish

1. Introduction

Although around 1.9 million people in Ireland report some level of proficiency in the Irish language, its everyday use remains limited. According to the most recent census in 2022 (Central Statistics Office, 2023), only about 72,000 people speak the language on a daily basis, despite recent increases in the overall number of speakers. Though over 39% of the population indicated some ability to speak Irish, English remains the dominant language used in most social and professional settings. Consequently, technological advancements – such as those in Automatic Speech Recognition (ASR) – have been sparce, with the exception of the seminal work of the Abair project (Chasaide et al., 2017).

However, as large public datasets at the scale of higher resource languages are not currently available, developments in ASR for Irish, as with many lower resource languages, remains challenging. This lack of fundamental technologies, specifically language technologies, has created a digital divide between Irish and English resources while also contributing to the potential for *digital extinction* of Irish. The results of this divide is speakers of the language reverting to using English, with only 1.5% of the population using the language outside of the education system (Lynn, 2023). This is in line with previous research classifying the Irish language as "definitely endangered" (Chiaráin et al., 2022).

Enormous strides have been made in ASR over the last decade with the advent of large, parallelisable deep learning models such as Whisper (Radford et al., 2023) and Wav2vec (Schneider et al., 2019), but much of these advancements have centred on ASR for higher resource languages such as English. This gap is well recognised (Li et al., 2022), leading to a variety of approaches to address the deficits of ASR for low-resource languages. For example, transfer learning is a commonly proposed approach to train models in scenarios where high quality transcribed data is unavailable, a common challenge in the field of ASR.

In this paper, we propose a more guided approach to transfer learning in which a pre-trained multi-lingual ASR model is first fine-tuned on one language, before being further fine-tuned on the target language – in this case Irish. We propose a language similarity-based approach to the selection of data used in transfer learning, and a more careful, multi-faceted analysis of downstream effects the data we use has on performance. Our objective is to disentangle the dataset characteristics that lead to an increase in performance from the aspects of learned representations that are still unknown. With the available open source data for Irish as a case study, we:

1. Demonstrate how language and dataset selection affect transfer learning performance.
2. Disentangle the characteristics of a model's

Proceedings of SPEAKABLE @ LREC 2026, pages 162–173
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

training dataset (e.g., dataset size, similarity with the target language) on the efficacy of transfer learning.
3. Determine whether these characteristics can be used to find the optimal pre-training dataset.
4. Measure the effects of different multi-step fine-tuned models on the performance across a variety of demographics and dialects of Irish.

The remainder of the paper is structured as follows: Section 2 provides background and discusses related work on transfer learning and representation learning for ASR and low-resource ASR. Section 3 details the proposed transfer learning method and the linguistic and dataset characteristics we focus on. Section 4 states the research questions that were investigated and describes the experimental setup used to evaluate them. Section 5 presents the results along with a detailed analysis of performance disparities across demographic lines. Section 6 concludes the paper.

2. Background & Related Work

A foundational concept of transfer learning for low-resource learning is the notion of learning generalised representations that transfer across languages. We detail the approaches to and challenges surrounding learning these representations (Section 2.1), the potential role of linguistic similarity to guide this process (Section 2.2), and the role of users in determining whether this transfer of representations might be successful (Section 2.3).

2.1. Building Generalisable Representations

There have been several attempts to learn invariant speech representations for ASR. These are loosely considered to be latent features that are universal to all languages or common across several. One such approach is to create robust or “generalised" speech representations which can be transferred to several speech tasks. Kawakami et al. (2020) aimed to learn robust multilingual representations of speech which transferred to phonologically diverse languages and, indeed, found that such developments led to a marked improvement in WER for several low-resource languages. One issue with approaches such as these, however, is that these “invariant" features are intrinsically opaque and scarcely explored in any detail. Though transferable representations are likely instrumental to accurate low resource speech recognition, the true universality of these features is rarely evaluated beyond a single score for a collection of low-resource languages. To this end, a more focused evaluation of pre-trained bilingual and trilingual models was performed by Lehečka et al. (2024). In contrast to Kawakami et al. (2020), the authors found that monolingual models outperformed bilingual and trilingual Wav2Vec models on oral history archives (even when controlling for dataset size). A similar result was found by Babu et al. (2022) where monolingual English models outperformed a similarly sized multilingual XLS-R model, a phenomenon the authors termed *capacity dilution*. The XLS-R model only reduced the WER upon significantly increasing the number of parameters.

Given the mixed results obtained using multilingual models, it remains challenging for practitioners to reliably build effective ASR models using them. Though it is domain invariant features that are likely to be learned using these models, very little attempt is made to test whether *definable* invariant features are being learned. It is therefore challenging to determine what is actually being transferred from a source model to a target, and whether these features are even desirable to transfer.

2.2. Linguistic Similarity

Linguistic similarity is explored less in the literature with more focus on the use of large uncurated datasets. However, the notion of capacity dilution described by Babu et al. (2022) suggests that, with a fixed capacity architecture, some language-specific features are likely to be lost in favour of models with increased language coverage. Additionally, large datasets are often very imbalanced with a disproportionate percentage of examples being from higher resourced languages (Ardila et al., 2020). To address these issues, work has been done to determine similarities between source and target languages which centers around the use of language identification or phonetic classifiers.

Early work on this topic by Zhang et al. (2014) used the posterior scores over source languages for target-language utterances to determine similarity. The bottleneck features of the source language with the highest posterior score were then used for pre-training. This work found that the closest languages appeared to provide the most benefit to training, stating that there was *“no data like similar data".* Thomas et al. (2016) took a similar approach by discriminating between phonemes of all source languages, combining the resultant phoneme scores for each language into a global language score. These scores were then used to create a language similarity matrix on which spectral clustering was applied to create language clusters. Multilingual feature frontends were then trained using the language groups identified within each cluster. Qian and Zhou (2022) extended this idea by extracting hidden representations of a language identification model and obtaining the similarity of a given utterance’s embedding to the average embedding of the target language. The similarity was then used to

weight the loss from a given utterance in order to favour some utterances within a multilingual dataset over others depending on their similarity during training. This technique also led to improvements over standard multilingual pre-training. Li et al. (2019) also saw improvements when corpus-level embeddings were used to select related corpora for training, a method that outperformed fine-tuned multilingual models. Though seemingly effective, techniques using language identification or those based on corpus-level embeddings require training of a discriminative model to obtain similarity scores and rely on the efficacy of the model. They also lack interpretability. They do not provide information as to *why* two languages have been found to be similar given that the similarity is entirely data-driven and learned features cannot be definitively interpreted.

Datasets using the same language with dialectal differences have also been explored. Yi et al. (2020) evaluated the efficacy of supervised pre-training using the Libri 100h (Panayotov et al., 2015) (English dataset) against supervised pre-training using HKUST (Liu et al., 2006) (Mandarin dataset) on the target test set, CALLHOME-MA (also a Mandarin dataset). The authors found that pre-training on HKUST, a "target-similar dataset" Yi et al. (2020) was more effective. They also found that using an abundance of multilingual data for self-supervised pre-training (Libri 1000h (Panayotov et al., 2015)) did lead to superior performance when evaluated on the same CALLHOME-MA dataset over the supervised model trained on HKUST. However this analysis did not further explore the potential reasons for the improved performance or detail the characteristics of the dataset. We are taking a more interpretable approach to finding dataset similarity, aiming to develop a method that identifies dataset characteristics needed to enhance performance.

2.3. Building Models that Serve People

A factor that is also rarely considered in the work described above is the use cases or demographic coverage of ASR systems. There are however a few exceptions, for example, Kawakami et al. (2020) specifically considered the use case of speakers in their model development and evaluation, Jimerson et al. (2023) investigated which model architectures work best for various low-resource languages, and Mitra et al. (2016); Le Ferrand et al. (2020); Littell et al. (2016) among others have dedicated work to specific endangered languages to overcome challenges they face. However, in contrast, a majority of approaches are not designed for specific languages and assume that the use of large, uncurated, datasets will ensure coverage across demographics, this still remains to be seen. For instance, Markl (2022) found that even English models, targeting one of the most well-resourced languages in ASR, perform worse for marginalised communities, including those who speak English as a second language or use specific dialects. Koenecke et al. (2020) also found that common ASR services demonstrated higher error rates for individuals using AAVE. Reitmaier et al. (2022) warned that low-resource ASR is in danger of being treated as an *"intellectual challenge"* with training datasets neither being collected with sufficient input from the communities that speak the languages, nor curated to adequately serve them. This makes interpretable and multifaceted model evaluation important, as understanding the shortcomings of existing benchmarks and models is instrumental in developing more equitable technology.

3. Methodology

This section explains the process of multi-step fine-tuning (Section 3.1), then defines the dataset-dependent (Section 3.2) and independent (Section 3.3) metrics used to determine similarity.

3.1. Multi-step fine-tuning

When performing transfer learning, a large-scale multilingual or mono-lingual dataset is first used to train an ASR model. For language-specific ASR models, models are either trained "from scratch" or, more commonly, fine-tuned from a multilingual model. Our objective is to determine whether certain language-specific models can lead to increased performance when further fine-tuned on a target language. That is, where a language-specific model has been built by fine-tuning a multilingual model on a single language dataset, we seek to fine-tune the model a second time on the target language. For the purposes of this work, we will term this strategy as *multi-step fine-tuning* and we will refer to the languages used in the first stage of fine-tuning as *source fine-tuning languages*. The language used during the final stage of fine-tuning is referred to as the *target language*. We seek to understand whether this strategy is more effective than fine-tuning a multilingual model on Irish alone, whether an increase in performance is uniform across different source fine-tuning languages, and whether the same multi-step fine-tuned model leads to the same performance gains across dialects and demographics.

3.2. Dataset Size

Dataset size is consistently associated with improved performance for deep learning architectures (Lehečka et al., 2024; Kawakami et al., 2020; Yi et al., 2020; Yusuyin et al., 2025). We evaluate if the size of the source fine-tuning dataset is

associated with the performance on the target language. While ASR datasets are typically described in terms of speech duration, we use number of samples since Common Voice clips have relatively consistent lengths.

3.3. Language Similarity/Proximity

We hypothesis that dataset-independent language features between the source fine-tuning languages and the target language also influence the model performance. We used the following two methods to calculate this similarity:

3.3.1. Genetic Proximity

Genetic Proximity is computed using a "Genetic Proximity Calculator" tool provided by eLinguistics.net (Elinguistics, 2020) and is independent of the datasets used to train models. This tool calculates the genetic proximity between two languages based on a cognate score (a metric to quantify the similarity between words across different languages). It is derived by comparing the consonants of 18 words that are commonly used in comparative linguistics studies. Consonants specifically are compared as they tend not to evolve as quickly as vowels (Vincent and Johannes, 2020). This tool outputs a score between 0 and 100 calculated using the cognate scoring and statistical context, indicating if the languages are similar or unrelated.

3.3.2. Averaged Lang2vec Similarity

Lang2vec is a Python tool used to query the URIEL database that represents languages using typological, phylogenetic, and geographical features such as genetic, geographic, syntactic, inventory, phonological, and featural vectors (Littell et al., 2017). It provides pre-calculated cosine distances based on these features that can be used to measure language similarity, meaning this metric is also independent of the datasets used to train models. However, analysis performed by Toossi et al. (2024) on the reproducibility of URIEL's language distances indicated that "31.24% of the languages in URIEL have no linguistic feature information", yet it still provides the distances for these languages. This problem inevitably impacts the reliability of this measure for low-resource languages which tend to have more missing values. To address these missing feature vector values, we propose the following method of calculating cosine similarity between the source fine-tuning language and the target language instead of using the pre-calculated cosine distances:

- Find all complete overlapping feature sets between the target and source fine-tuning languages.
- Calculate the cosine similarity for each of these complete feature sets.
- The averaged lang2vec similarity score is the mean cosine similarity for each feature set.

This strategy of similarity score averaging means that all feature sets are given equal weighting and the scores are not influenced by the size of the sets.

4. Experimental Setup

This section includes the specific models and dataset used to carry out the experiments. We choose Irish to the be target language and Dutch, French, German, Persian, Portuguese and English to be the source fine-tuning languages. The research questions we explore are:

RQ.1 Is multi-step fine-tuning more effective than directly fine-tuning a multilingual model?
RQ.2 Is the performance increase provided by multi-step fine-tuning uniform across source fine-tuning languages?
RQ.3 Is higher proximity between the source fine-tuning language and the target language associated with better model performance?
RQ.4 Is the size of the source fine-tuning dataset associated with better model performance?
RQ.5 Is multi-step fine-tuned model performances uniform across dialects and demographics?

4.1. Data

With Irish as the target language for our experiments, we used the open source Common Voice dataset version 15.0 (Mozilla Corporation, 2021) for Irish fine-tuning with its training (536 clips), validation (516 clips), and testing[1] (517 clips) splits. Special characters were removed, words were converted to lowercase and a vocabulary was built using tokenisation. Audio was sampled at 16kHz and zero-padded to match the longest segment.

4.2. Model

The base model in all cases was the Wav2Vec2 cross-lingual speech representation large-scale model trained on 53 languages (XLSR-53) (Conneau et al., 2020). The "feature encoder" of each base model was frozen in the initial fine-tuning stage. Then, this was done again in our experiments using the Irish data. This means that the transformer encoder or "context network" is

[1] Conneau et al. (2020) states all Common Voice languages were used during unsupervised pre-training of XLSR-53 however it is unclear which training/test splits were used for training in the original paper. Despite this, the Irish portion of the dataset has since expanded compared to what was available at the time of this pre-training.

fine-tuned with CTC and the lower level "feature encoder" is frozen to preserve speech features learned from pre-training. Each model configuration was fine-tuned in three independent runs, and the reported results represent the average performance across these runs.

4.2.1. Baseline Single-Step Fine-Tuning

The baseline for experiments was single-step fine-tuning using Irish. Hyperparameters were selected using Bayesian hyperparameter optimisation (Yang and Shami, 2020) with Word Error Rate (WER) as the objective. The best configuration was achieved using a learning rate of 0.0003, a batch size of 8, and 45 training epochs. We note that the objective here is not to create the optimal model for Irish ASR but to find reasonable hyperparameters that can be used uniformly across models in order to compare them.

4.2.2. Multi-Step Fine-Tuning

For our source fine-tuned models, we use XLSR-53 models that are openly available and already fine-tuned for our choice of source fine-tuning languages provided by Grosman (2021). These models had already completed the first stage in the multi-step fine-tuning method using the train and validation splits of Common Voice 6.1 (along with additional audio clips from CSS10 (Park and Mulc, 2019) for Dutch only). To create the multi-step fine-tuned models, they are fine-tuned again on Irish using the same pre-processing steps, data, and hyperparameters as the baseline described in Section 4.2.1.

4.3. Averaged Lang2vec Similarity

As discussed in Section 3.3.2, some source fine-tuning languages chosen for our experiments didn't contain the same complete feature sets as Irish, the target language. French, German, and Persian contained complete sets of the same features. English, Portuguese, and Dutch were missing values that were present for Irish, therefore these feature sets were not included when calculating the averaged lang2vec similarity for those respective languages.

4.4. Demographic and Dialects

We sought to determine whether models performed the same for different speaker demographics and dialects. To do this, we filtered the unseen Irish test dataset to isolate data from each demographic group annotated within the Common Voice dataset (e.g, females, males, teens, 20s, etc.).

Table 1: WER(%) of the different Models

Model	WER	# training samples
English	56.7	580501
French	57.6	314745
Portuguese	59.0	11106
Persian	61.4	12806
Dutch	61.4	14398
Irish (baseline)	66.2	536
German	94.9	262113

5. Results

Section 5.1 details the results of multi-step fine-tuning by comparing the performance of the different models. The results concerning the effect of language and dataset characteristics are split across three sections. First, Section 5.2 analyses the degree to which performance is associated with dataset training size. Second, Section 5.3 focuses on the relation between performance and the dataset-independent language similarity metrics. Finally, Section 5.4 provides an analysis of the performance difference along demographic lines.

5.1. Multi-step Fine-tuning Comparison

Table 1 shows the models' WER performance and source fine-tuning dataset size. Interestingly, we see the benefit provided by multi-step fine-tuning is dataset-specific. In most cases there is an improvement, with the exception of German, where we see a large decrease in performance with a relative percentage increase in WER of $\sim 43\%$. English provides the most improvement with a relative percentage decrease in WER of $\sim 14.4\%$, followed by French and Portuguese with a decrease of $\sim 13\%$ and $\sim 10.9\%$ respectively. This motivates our conclusion to **RQ.1** and **RQ.2**, that multi-step fine-tuning provides a benefit and that the specific source fine-tuning language used impacts performance.

5.2. Effect of Dataset Size

Figure 1 plots the number of training samples from each source fine-tuning dataset against their models' WER performance. This shows that the number of training samples used in the first step of fine-tuning is not clearly associated with increased performance in WER. A Pearson correlation coefficient of 0.008 between WER and the size of the datasets indicates no strong linear relationship. While the English model is fine-tuned on the largest dataset and achieves the lowest WER rate, the French model is trained on a dataset almost half

Table 2: Ranked Similarity Scores of the Different Datasets in Comparison to the Irish Dataset

(a) Genetic proximity between languages

Dataset	Genetic Proximity
French	57.7
Portuguese	59.7
Persian	60.6
German	76.5
English	78.5
Dutch	80.8

(b) Averaged Lang2vec similarity

Dataset	Averaged lang2vec
English	0.6842
Portuguese	0.6558
Dutch	0.6076
Persian	0.5938
German	0.5923
French	0.5922

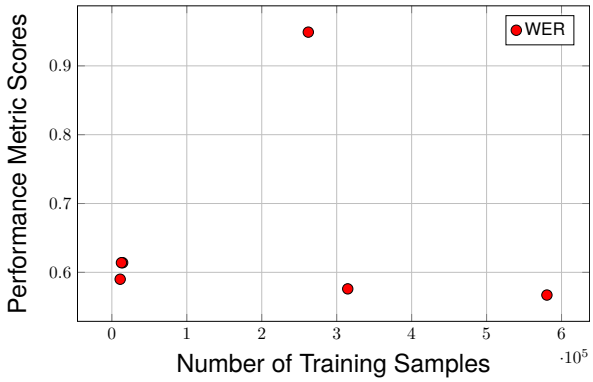


Figure 1: Relationship between the Number of Training Samples and Performance Metrics

the size of English, but there is only a relative percentage difference of $\sim 1.6\%$ in WER. Looking at the model first fine-tuned on Portuguese, it uses a dataset $\sim 1.9\%$ the size of English, and there is only a relative percentage increase of $\sim 4\%$ in WER. The German dataset is the third largest dataset and it achieves the worst performance. This appears to show that the number of examples alone cannot explain the increase in performance against the baseline, answering **RQ.4**, and that the dataset size *in conjunction with* other factors should be considered.

5.3. Language Similarity/Proximity Comparison

When using more generalised similarity metrics such as averaged Lang2vec and Genetic Proximity, we observe some strong correlations between performance and similarity. Specifically, Genetic Proximity of languages in Table 2a has a positive Pearson correlation co- efficient of 0.347. This could indicate phonetic similarities between languages plays an important role in boosting performance, but analysis across additional languages is necessary to make this claim more robust. Interestingly, the correlation between the averaged Lang2vec similarity metric shown in Table 2b and the performance of the models is negatively associated with a correlation of -0.4253. This could be due to features irrelevant to speech being considered. This motivates future further analysis on feature selection.

5.4. Performance Variation across Demographic Lines

While WER offers a general performance overview, we observed that it varied across demographics. We compared the WER by gender (Section 5.4.1), age (Section 5.4.2), and dialect (Section 5.4.3).

5.4.1. Gender

The annotated gender categories in the unseen Irish test data of common voice were: `Female`, `Male` and `Other`. When all the models were evaluated on this dataset, the models first fine-tuned on Portuguese (Figure 2e), German (Figure 2f), Persian (Figure 2d) and Dutch (Figure 2c) as well as the baseline model only fine-tuned on Irish (Figure 2g), performed worse for the `Female` category compared to `Male`. While, on face value, this performance disparity seems due to the lack of `Female` representation in datasets, it should be noted that the annotations are incomplete. There were many utterances in all datasets without gender labels. Of all gender labels in the metadata of the training and validation splits, only 11% were labelled `Female` in the Irish dataset, 10.7% in the German dataset, 9.4% in the Dutch dataset, 19.8% in the Persian dataset and as little as 5.8% in the Portuguese dataset. In all cases the `Other` gender category was the most under-represented. With no `Other` gender category labels in the Irish test dataset, we were unable to see the effect this under-representation had on the performance.

Interestingly, for models with English (Figure 2a) and French (Figure 2b) as the source fine-tuning language, audio labelled `Female` is in general more accurately detected than those labelled `Male`. Of the existing gender annotations, the French dataset has only 13.9% `Female` samples (Figure 3b) and the English dataset has only 25.5%(Figure 3a).

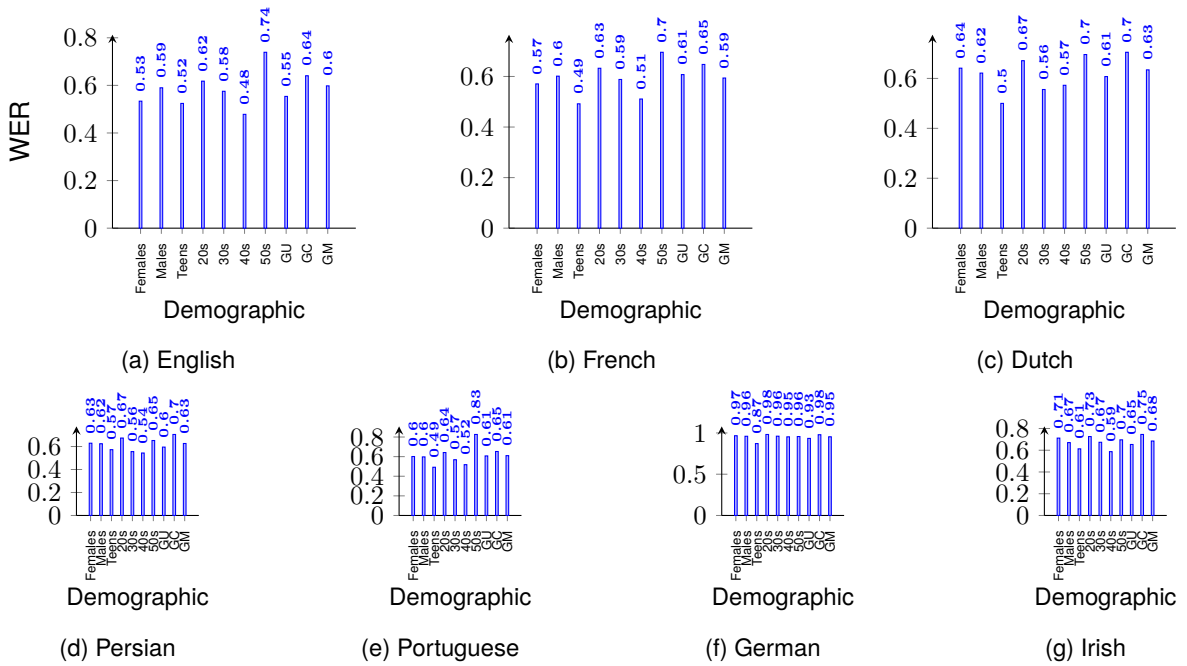


Figure 2: WER of each Model by Demographic

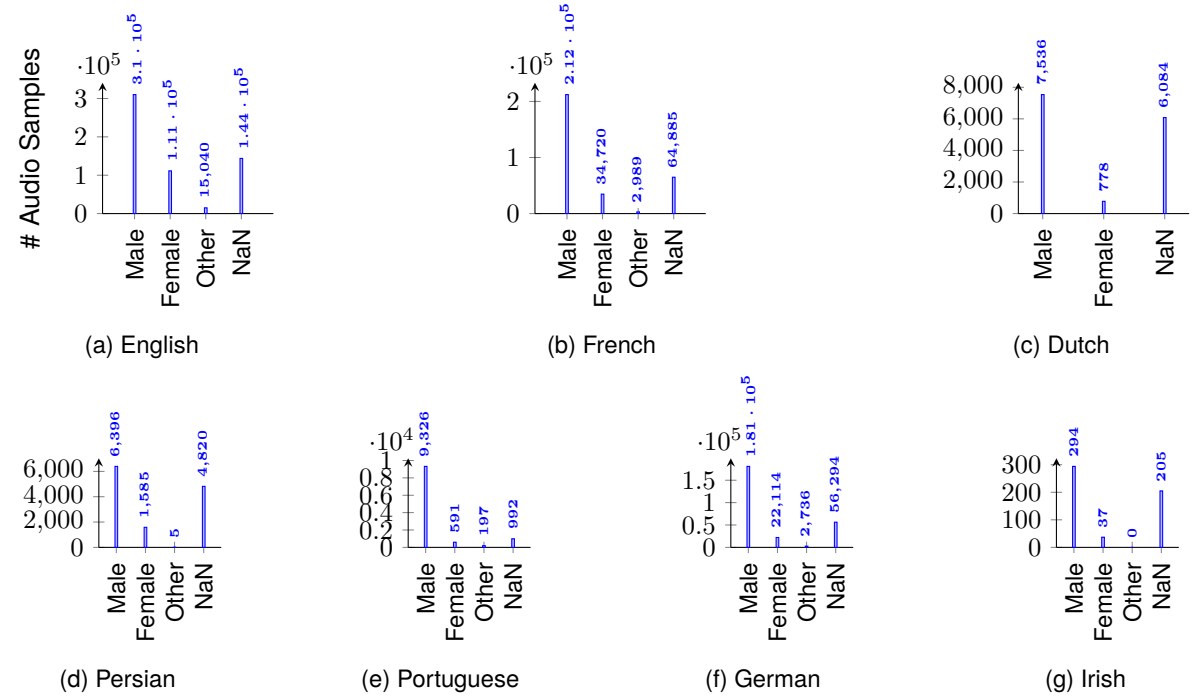


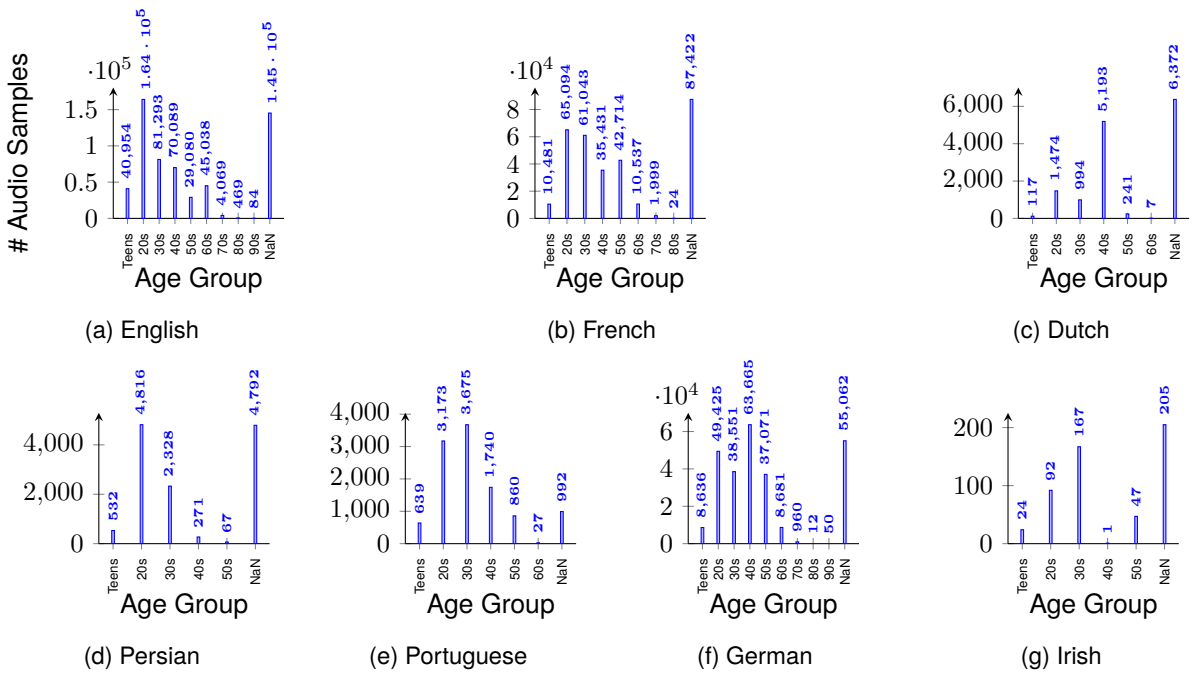
Figure 3: Gender Distribution of Training Datasets

Due to the missing gender annotations, it is challenging to deduce whether performance disparity between genders is due to dataset imbalance, however it is clear that such a disparity exists.

5.4.2. Age

In Figure 2 we observed a notable variation in performance across age-groups. This aligns with findings by Werner et al. (2019) showing ASR systems being impacted by the age of speakers. Moore (2011) also discusses variations in pronunciation unique to younger speakers, which seems to be reflected in our results since in general audio labelled `Teenager` performs well across all models.

Interestingly, in the French, Portuguese, Dutch, and German datasets (Figure 4), teenagers are the smallest represented age group, *based on the*



(a) English (b) French (c) Dutch

(d) Persian (e) Portuguese (f) German (g) Irish

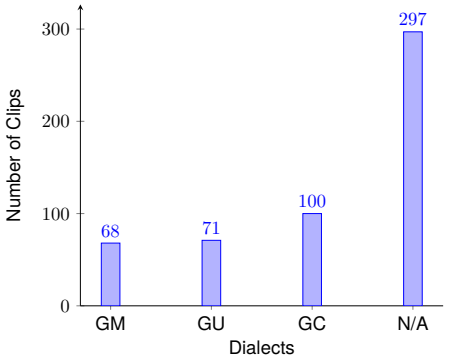
Figure 4: Age Distribution of Training Datasets

annotations, out of all age groups in the test set.

5.4.3. Dialect

We now move to discuss the performance variation across the three dialects labelled within the Irish common voice dataset: Gaeilge Uladh (GU), Gaeilge Chonnacht (GC) and Gaeilge na Mumhan (GM). Apart from the model source fine-tuned on French, audio labelled GU had the lowest WER. The GU dialect has the second most labelled audio associated with it, but GM only falls slightly behind by three audio clips (Figure 5). However, with a significant number of missing dialect labels, it remains unclear if the number of training utterances is a contributing factor in this bias towards GU. The performance difference is, perhaps, better explained by the fact that GU is more distinct from the other two dialects, which models have apparent difficulty disambiguating between. Lonergan et al. (2023a) observed similar behaviour for dialect classification.

Assuming that the distribution of the labelled dialects accurately reflects the true distribution within the data, we can see that the balance of the dialect classes do not necessarily lead to proportionally balanced WER. For example, GC is the most represented dialect but consistently performs worst. This could align with findings by Lonergan et al. (2023a) that balanced corpora do not necessarily lead to balanced performance. Further investigation is needed to determine the sources of this bias.



GM: Gaeilge na Mumhan, GU: Gaeilge Uladh, GC: Gaeilge Chonnacht, N/A: Unknown

Figure 5: Dialect Distribution in Irish Training Dataset

6. Conclusion

This work has demonstrated the advantages and remaining challenges for ASR in Irish using open-source datasets. We demonstrated that, though multi-step fine-tuning can provide performance gains over single-step fine tuning in many cases, these gains are not uniform across different source fine-tuning languages, and more specifically are often not uniform across demographic lines. Therefore, a one-size-fits-all approach is not effective, highlighting the need for detailed analysis of the model's performance instead of a single evaluation metric.

While the usefulness of a dataset often gets overlooked in Machine Learning, we attempted to un-

derstand what dataset performs better for a certain low-resource language. We defined this usefulness using both dataset-dependent and independent similarity measures. From our case study of six source languages, we show that certain measures, in particular Genetic Proximity, can be indicative of the optimal pre-training model. While increasing this study to more languages could establish these measures as robust predictors, from our experiments it seems these metrics alone can't show which model will work best, meaning that to tackle this problem we need to consider more factors.

Interestingly, through our experiments on dataset-dependent features, we found that dataset size is not strongly associated with performance gains, indicating that blindly increasing dataset size is unlikely to result in improved performance. Furthermore, the percentage increase in data points does not translate to proportionate performance gains.

We also found that there are often gaps in demographic information. Given that speech data can often identify the speaker, collecting it may be too high-risk, especially for speakers of marginalised groups. This motivates a privacy preserving alternative for measuring demographic coverage.

7. Ethics/Broader Impact Statement

This work relies exclusively on open-source, publicly available datasets for the purposes of reproducibility. Our work focuses on effective model performance in a low-resource setting in order to make these type of speech recognition systems more accessible for different communities. We investigated our method for one specific language (Irish) in order to promote research into the nuances and the unique characteristics of each individual language, something that often gets overlooked in the development of speech recognition models. We evaluate our models' performance across different demographics in order to assess the biases that exists within these systems and to highlight that a single aggregate metric does not take into account the variation in model performance for different speakers.

In our limitations section, we outline that a large amount of speaker metadata is missing from the datasets used in experiments. Though additional metadata labels would undoubtedly facilitate a more rigorous analysis of performance differences between, and at the intersection of, different demographic groups, it is also important for those curating datasets and researchers evaluating ASR to study and understand *why* participants do not feel comfortable disclosing personal information (age, gender etc.) while contributing to datasets. More specifically, it is incumbent on the community to develop ways of evaluating ASR that both protect data subjects and their privacy, while also rigorously testing model errors through an intersectional lens. Based on the lack of speaker metadata in the datasets studied, we endeavor to research this aspect of evaluation further in future work.

8. Limitations

Only open-sourced data from the Common Voice dataset was used in these experiments. Using more datasets could improve the reliability of our experiments. We also only considered six source fine-tuning languages and the base model in all cases was Wav2Vec. Scaling this experiment to more language datasets and base models (such as OpenAI's newer Whisper model (Radford et al., 2023)) could have increased the robustness of our findings and allowed us to detect other factors that lead to performance gains. The source fine-tuned models provided by (Grosman, 2021) do not perfectly match our training conditions. This be another factor contributing to the impact the source-language has on the performance on the model. Another limitation of using the Common Voice was that some language datasets were not available. These included Scottish Gaelic and Manx, which are in the same the Celtic language family as Irish. As described throughout the paper, lack of dataset availability is a common issue within ASR research for low-resource languages such as these.

A large portion of the datasets in Common Voice are missing annotations. This had implications for our study as we endeavored to investigate the link between performance and demographic representation in the source fine-tuning dataset. Certain demographic groups in the datasets were missing more labels than others. For example, of the gender labels, `Female` and `Other` were available for considerably fewer utterances than `Male`. This either indicates that those who do not identify as male are reluctant to disclose their gender, or that they are under-represented in the dataset in the first place. Similarly to the gender labels, age and dialect labels were also missing. It was difficult for us to draw conclusions about the performance across different demographics given this incomplete information. A broader sociolinguistic analysis of utterances is likely necessary to determine the reasons for disparate performance across examples. Furthermore, we also note that we have not analysed the intersection between demographics, which are themselves likely to reveal differences in performance. It is clear that more detailed analysis of pre-training datasets and models is necessary to disentangle the sources of performance differences.

9. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2278–2282. ISCA.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Central Statistics Office. 2023. Press statement census 2022 results profile 8 - the irish language and education.

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl. 2017. The abair initiative: Bringing spoken irish into the digital space. In *INTERSPEECH*, pages 2113–2117.

Neasa Ní Chiaráin, Oisín Nolan, Madeleine Comtois, Neimhin Robinson Gunning, Harald Berthelsen, and Ailbhe Ní Chasaide. 2022. Using speech and nlp resources to build an icall platform for a minority language, the story of an scéalaí, the irish experience to date. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 109–118.

Ruth Holmes, Ellen Rushe, Mathieu De Coster, Maxim Bonnaerens, Shin'ichi Satoh, Akihiro Sugimoto, and Anthony Ventresque. 2023. From scarcity to understanding: Transfer learning for the extremely low resource irish sign language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2008–2017.

Robert Jimerson, Zoey Liu, and Emily Prud'Hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. 2020. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1182–1192. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428.

Jan Lehečka, Josef V Psutka, Luboš Šmídl, Pavel Ircing, and Josef Psutka. 2024. A comparative analysis of bilingual and trilingual wav2vec models for automatic speech recognition in multilingual oral history archives. In *Interspeech 2024*, pages 1285–1289.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Xinjian Li, Siddharth Dalmia, Alan W. Black, and Florian Metze. 2019. Multilingual speech recognition with corpus relatedness sampling. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2120–2124. ISCA.

Patrick Littell, Kartik Goyal, David R Mortensen, Alexa N Little, Chris Dyer, and Lori Levin. 2016. Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 998–1006.

Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide.

2022. Automatic speech recognition for irish: the abair-éist system. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023a. Towards spoken dialect identification of irish. In *Proceedings of the 2nd annual meeting of the Special Interest Group of Under-resourced Languages, a Workshop at Interspeech 2023*.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023b. Towards dialect-inclusive recognition in a low-resource language: Are balanced corpora the answer? In *Proc. Interspeech 2023*, pages 5082–5086.

Teresa Lynn. 2023. Language report irish. In *European Language Equality: A Strategic Agenda for Digital Language Equality*, pages 163–166. Springer.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 521–534.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Robert Moore. 2011. "if i actually talked like that, i'd pull a gun on myself": accent, avoidance, and moral panic in irish english. *Anthropological Quarterly*, pages 41–64.

Yanmin Qian and Zhikai Zhou. 2022. Optimizing data usage for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:394–403.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–17.

Samuel Thomas, Kartik Audhkhasi, Jia Cui, Brian Kingsbury, and Bhuvana Ramabhadran. 2016. Multilingual data selection for low resource speech recognition. In *Interspeech*, pages 3853–3857.

Hasti Toossi, Guo Qing Huai, Jinyu Liu, Eric Khiu, A Seza Doğruöz, and En-Shiun Annie Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the uriel knowledge base. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, NAACL 2024, Mexico City, Mexico, June 18, 2024*, pages 233–241. Association for Computational Linguistics.

Beaufils Vincent and Tomin Johannes. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration.

Lauren Werner, Gaojian Huang, and Brandon J Pitts. 2019. Automated speech recognition systems and older adults: a literature review and synthesis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 42–46. SAGE Publications Sage CA: Los Angeles, CA.

Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, and Alan W Black. 2021. Cross-lingual transfer for speech processing using acoustic language similarity. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1050–1057. IEEE.

Li Yang and Abdallah Shami. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Saierdaer Yusuyin, Te Ma, Hao Huang, Wenbo Zhao, and Zhijian Ou. 2025. Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. *IEEE Transactions on Audio, Speech and Language Processing*.

Yu Zhang, Ekapol Chuangsuwanich, and James Glass. 2014. Language id-based training of multilingual stacked bottleneck features. In *Proc. Interspeech*, pages 1–5. Citeseer.

10. Language Resource References

Conneau, Alexis and Baevski, Alexei and Collobert, Ronan and Mohamed, Abdelrahman

and Auli, Michael. 2020. *Unsupervised cross-lingual representation learning for speech recognition*. Model available on HuggingFace: https://huggingface.co/facebook/wav2vec2-large-xlsr-53.

Elinguistics. 2020. *Elinguistics*. Accessed: 2025-02-10.

Grosman, Jonatas. 2021. *Fine-tuned XLSR-53 large models for speech recognition*.

Littell, Patrick and Mortensen, David R and Lin, Ke and Kairis, Katherine and Turner, Carlisle and Levin, Lori. 2017. *URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors*.

Liu, Yi and Fung, Pascale and Yang, Yongsheng and Cieri, Christopher and Huang, Shudong and Graff, David. 2006. *Hkust/mts: A very large scale mandarin telephone speech corpus*. Springer.

Mozilla Corporation. 2021. *Mozilla Common Voice*. Accessed: 29th September 2023.

Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev. 2015. *Librispeech: an asr corpus based on public domain audio books*. IEEE.

Park, Kyubyong and Mulc, Thomas. 2019. *CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages*.

Radford, Alec and Kim, Jong Wook and Xu, Tao and Brockman, Greg and McLeavey, Christine and Sutskever, Ilya. 2023. *Robust speech recognition via large-scale weak supervision*. PMLR.

Schneider, Steffen and Baevski, Alexei and Collobert, Ronan and Auli, Michael. 2019. *wav2vec: Unsupervised pre-training for speech recognition*.