university of groningen / campus fryslân

Assessing Knowledge-Distillation Based Compression of Whisper Model for Frisian ASR

Yanpei Ouyang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Assessing Knowledge-Distillation Based Compression of Whisper Model for Frisian ASR

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Shekhar Nayak (Voice Technology, University of Groningen)
with the second reader being
Dr. Phat Do (Voice Technology, University of Groningen)

Yanpei Ouyang (S5726956)

June 11, 2024

Acknowledgements

First and foremost, I want to extend my heartfelt thanks to the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster. Without their computational power, I'd still be running my first simulation. You guys rock!

I am deeply grateful to my supervisor, Shekhar, for his invaluable guidance and support. His wisdom and patience have seen me through some pretty tough times. Additionally, I also want to express my sincere appreciation to Matt, his advice on my abstract was very useful.

Furthermore, I would like to convey my immense gratitude to my parents and friends, their unwavering emotional support has been my rock! Here a special thanks to my friend, Xueying, for being my coding partner. We have conquered countless bugs and celebrated each successful run with an embarrassing dance-off. Her companionship has been nothing short of legendary.

I also want to thank my amazing classmates for their camaraderie and support. Whether it was through study sessions, shared resources, or simply being there to talk things through, you all have made this journey much more enjoyable and manageable.

I cannot forget to thank myself for the perseverance, hard work, and countless hours spent on this thesis. A special mention to my loyal computer. Together, we've created a masterpiece (or so I like to think). Your steadfast performance has been crucial in getting this thesis done.

Lastly, a general thank you to anyone and anything that contributed to the completion of this thesis. Whether you offered a word of encouragement, a cup of coffee, or simply a moment of distraction, you have my gratitude.

Cheers to all the wonderful people and the quirky moments that made this journey memorable!

Abstract

Multilingual ASR systems face challenges in accommodating diverse linguistic landscapes, particularly for low-resource languages (LRLs) with limited data. This study investigates the efficacy of model compression techniques, specifically knowledge distillation (KD) and fine-tuning, in enhancing the performance and efficiency of the Whisper-small model for LRLs. The research aims to determine whether applying KD and fine-tuning to the Whisper-small model can improve its performance on LRLs while reducing its computational and memory requirements. Fine-tuning experiments were conducted on both the English (LibriSpeech) and Frisian (CommonVoice 6.1) datasets for both the original Whisper-small model and the distilled Whisper-small model. Subsequently, a comprehensive evaluation based on various metrics, including Word Error Rate (WER), number of model parameters, and training set sizes, was performed. The results demonstrate that the distilled Whisper-small model achieved a WER of 26.91% when fine-tuned with 10 hours of Frisian data, exceeding the initial reduction target. In comparison, the Whisper-small model achieved a WER of 22.42% under the same conditions. Additionally, the distilled model showed competitive performance with limited training data, highlighting the potential of KD to create efficient ASR models suitable for environments with constrained computational resources and data availability. Furthermore, while the Whisper-small model supports recognition of many languages, including Dutch, it was successfully fine-tuned to recognize Frisian, a language it originally did not support. Similarly, the Distil-Whisper-small model, which initially only supported English, was also successfully adapted to recognize Frisian, showcasing the adaptability of these models for cross-linguistic applications. In conclusion, the findings validate the effectiveness of model compression techniques, particularly KD, in enhancing the performance and efficiency of ASR models for LRLs. This study contributes to the development of more efficient and inclusive multilingual ASR systems, providing valuable insights into optimizing ASR models for diverse linguistic landscapes, especially those with limited datasets. The implications of this research extend to various domains, including education, healthcare, and accessibility, ultimately advancing universal accessibility and real-world applications of ASR technology.

Keywords: Automatic Speech Recognition (ASR), fine-tuning, knowledge distillation (KD), Whisper-small Model, Frisian Language.

Contents

1 Introduction

ASR technology has transcended traditional boundaries, embedding itself across a spectrum of applications from virtual assistants to real-time captioning services. As smart devices become ubiquitous, the imperative for ASR systems to operate within the computational and memory constraints of these devices has never been more pronounced. In response to these challenges, model compression techniques, such as knowledge distillation, pruning, and quantization, have gained prominence. These methods aim to refine ASR models, including Whisper, to create versions that are both compact and efficient, facilitating their deployment on edge devices. This quest for optimization has not only preserved but, in many instances, enhanced the models' performance, making ASR technology more accessible and versatile.

The complexity of accommodating the world's linguistic diversity within ASR technology introduces additional challenges, particularly for LRLs. Despite advancements in model compression, as demonstrated by FFerraz, Boito, Brun, and Nikoulina (2024) and Yang et al. (2023), research on the performance of compressed models in recognizing languages with limited data, such as Frisian, remains sparse. These languages, rich in their linguistic uniqueness but constrained by data availability, underscore the need for an in-depth exploration of ASR model efficacy in a multilingual context. This thesis seeks to address this gap by conducting a comparative analysis of the fine-tuning of the distil-whisper-small model with Frisian data against the baseline Whisper-small model. By dissecting the capabilities of these models in handling the intricacies of such low-resource languages, this study endeavors to not only benchmark their performance but also to identify pathways for enhancing Whisper's utility across the diverse linguistic landscape of our world.

Note: Throughout this thesis, the terms "distil-whisper-small," "distilled Whisper-small," and "distil-small" refer specifically to the distil-small.en version of the Distil-Whisper model.

1.1 Research Question and Hypothesis

Based on the preceding discussion, the central research question of this study can be framed as follows:

> **How does fine-tuning the distil-whisper-small model with Frisian data affect its ASR performance compared with the baseline model (Whisper-small), considering Word Error Rate (WER), number of model parameters, training set sizes, and number of trainable parameters, across both high-resource (English) and low-resource (Frisian) languages?**

This primary question leads to several subquestions:

- What is the baseline WER achieved by fine-tuning Whisper-small on Frisian speech?
- Can the distil-whisper-small model achieve a lower WER than Whisper-small on Frisian speech?
- How does the performance of Whisper-small and distil-whisper-small models vary with different sizes of training data on Frisian speech, particularly in low-resource settings?

My hypothesis is that the distil-whisper-small model will achieve performance comparable to the Whisper-small baseline model when fine-tuned on Frisian speech, specifically aiming for a significant reduction in WER, potentially around 30%. It is anticipated that the size of the training data will substantially impact the model's performance, with larger datasets leading to better results. However, it is hypothesized that the distil-whisper-small model, when fine-tuned on a larger dataset (e.g., 1-hour Frisian dataset), will achieve similar performance to the Whisper-small model fine-tuned on a smaller dataset (e.g., 10-minute Frisian dataset). This reflects the effectiveness of model distillation in enhancing efficiency and maintaining performance. Additionally, due to its reduced number of parameters, the distil-whisper-small model is expected to offer faster speech inference and reduced computational requirements for both training and inference compared to the Whisper-small model. This should result in a favorable balance between training and inference time and overall ASR performance, making the distil-whisper-small model more suitable for deployment in resource-constrained environments.

2 Literature Review

The advent of ASR technologies like Whisper has significantly broadened the horizons of human-computer interaction, enabling applications that range from virtual assistants to real-time transcription services. Despite substantial advancements, the deployment of ASR systems that effectively handle multilingual content, especially in languages with limited resources, presents ongoing challenges. This section explores the development and optimization of ASR technologies with a focus on model compression techniques, particularly knowledge distillation. Additionally, it examines the specific challenges faced by LRLs and the performance metrics used to evaluate ASR systems.

The evolution of ASR technologies has been paralleled by efforts to optimize these systems for performance within the computational confines of modern devices. Techniques such as knowledge distillation, pruning, and quantization have been pivotal. Knowledge distillation, in particular, has shown promise in retaining model performance while significantly reducing computational demands (Sanh, Debut, Chaumond, and Wolf (2019)). Sanh et al. (2019) effectively utilized KD to create DistilBERT, achieving significant reductions in model size and increases in speed, while retaining a high level of performance on the GLUE benchmark. This principle was further extended to sequence-to-sequence models by Shleifer and Rush (2020), who developed DistilBART, demonstrating that KD could be effectively applied beyond the realm of encoder-only models, to more complex architectures with notable success in both model compression and speed improvements.

The Whisper model by OpenAI represents a leap forward in multilingual ASR, designed to comprehend a wide array of languages. However, studies like those by M. Shao, Li, Peng, and Sun (2023) have begun to explore its compression through knowledge distillation, aiming to maintain, or even enhance, its linguistic versatility within a more compact framework. These endeavors underscore the potential for high-performance, efficient ASR models that are accessible on edge devices. Low-resource languages, such as Frisian, pose unique challenges due to the scarcity of data and their complex linguistic features. Despite the inclusion of such languages in large-scale multilingual models, detailed analyses of ASR performance, specifically under compression, are sparse.

Performance evaluation of ASR systems traditionally revolves around metrics like WER, which directly quantifies transcription accuracy. Additional metrics, such as the number of model parameters, play crucial roles in assessing the practical deployment of ASR models in real-world settings. The application of these metrics in prior research offers a foundation for comprehensive performance analysis, particularly in assessing the adaptability of ASR technologies to diverse and challenging linguistic landscapes. This thesis aims to bridge the research gap by conducting an in-depth evaluation of the fine-tuning of the distil-whisper-small model with Frisian data against the baseline Whisper-small model. The literature review is organized into five sections. Subsection 2.1 provides a historical context and outlines the fundamental principles and advancements in ASR technology. Subsection 2.2 discusses the unique challenges in developing ASR systems for LRLs, including data scarcity and linguistic diversity, and presents detailed case studies illustrating practical applications and the effectiveness of these techniques in LRLs. Subsection 2.3 describes various model compression techniques, including knowledge distillation, fine-tuning, pruning, and quantization, and their applications in ASR. Subsection 2.4 reviews studies on how these compression techniques enhance the efficiency and performance of ASR systems. Subsection 2.5 discusses the metrics used to evaluate ASR system performance, such as WER and model parameters. Furthermore, I identify the limitations of current research and suggests potential future research directions in the end.

Through this structured exploration, the literature review aims to provide a comprehensive under-

standing of the current state of ASR technologies, the challenges of multilingual and low-resource language processing, and the potential avenues for future advancements.

2.1 Overview of ASR Technology

ASR technology has significantly advanced over the past few decades, transforming from simple command-based systems to complex models capable of understanding and processing natural language. The primary goal of ASR is to convert spoken language into text accurately and efficiently. Early ASR systems relied heavily on handcrafted features and rule-based approaches, which limited their adaptability and scalability. With the advent of deep learning, particularly end-to-end (E2E) models, ASR systems have achieved remarkable improvements in accuracy and robustness. End-to-end models simplify the ASR pipeline by integrating acoustic, language, and pronunciation models into a single neural network framework, as demonstrated by Graves, Mohamed, and Hinton (2013), who showcased the efficacy of deep recurrent neural networks (RNNs) in speech recognition.

In multilingual environments, ASR systems face additional challenges due to the diversity of languages and dialects. This diversity necessitates the development of robust models that can generalize well across different languages. Zhou, Wang, Liu, Yu, and Xiang (2021) introduced a configurable multilingual model that can be trained once and configured as different language-specific recognizers, significantly reducing the complexity and cost associated with multilingual ASR systems. Furthermore, Babu et al. (2021) highlighted the importance of large-scale self-supervised cross-lingual speech representation learning, which has proven effective in enhancing ASR performance across both high-resource and low-resource languages.

2.2 Challenges of Low-Resource Languages

Developing ASR systems for LRLs presents unique challenges, primarily due to the scarcity of annotated data and the linguistic diversity among these languages. LRLs often lack the extensive linguistic resources and large annotated datasets available for high-resource languages, making it difficult to train robust ASR models. This scarcity of data leads to high error rates and poor model generalization, as highlighted by Yadav and Sitaram (2022), who explored the current state of multilingual ASR and the challenges associated with handling LRLs.

Addressing these challenges requires innovative approaches to data augmentation, transfer learning, and model optimization. Tsoukala, Lange, and Wang (2023) presented a practical implementation of an ASR pipeline for a low-resource language, detailing the challenges and solutions in data processing and model training. Their work emphasizes the importance of tailored data processing techniques and the adaptation of existing models to the specific characteristics of LRLs. Additionally, Khare, Mane, Singh, and Agarwal (2021) demonstrated the surprising effectiveness of pre-trained models in low-resource ASR scenarios, showing that leveraging large-scale pre-trained models can significantly enhance recognition performance with minimal additional data.

Overcoming the obstacles associated with LRLs in ASR systems requires detailed case studies that illustrate practical applications and the effectiveness of various techniques. One such case study is presented by Tsoukala et al. (2023), who implemented an ASR pipeline for the Pomak language, a low-resource language. They detailed the specific challenges encountered in data processing and model training, and the solutions they developed, which included tailored data augmentation techniques and the adaptation of existing ASR models to the unique characteristics of the language.

Farooq, Imran, and Pasha (2023) conducted a study on the effectiveness of the MUST learning approach, a multilingual student-teacher framework designed to address the scarcity of annotated data in LRLs. Their approach significantly reduced the character error rate in low-resource settings by leveraging the knowledge from high-resource languages. Additionally, Khare et al. (2021) highlighted the potential of pre-trained models in improving ASR performance for LRLs. By utilizing large-scale pre-trained models, they demonstrated significant enhancements in recognition accuracy with minimal additional data, underscoring the practicality and efficiency of transfer learning methods in low-resource scenarios.

2.3 Model Compression Techniques

Model compression techniques, such as KD and fine-tuning, play a crucial role in enhancing the efficiency and performance of ASR systems, particularly for deployment in resource-constrained environments. Knowledge distillation involves transferring the knowledge from a large, cumbersome model (teacher) to a smaller, more efficient model (student), which retains much of the teacher model's performance but with significantly reduced computational requirements. Hinton, Vinyals, and Dean (2015) described the process of knowledge distillation, highlighting its effectiveness in maintaining model performance while reducing complexity. Xu, Liu, and Chang (2024) further explored various KD techniques applied to large language models, providing insights into the implementation and efficacy of KD in improving ASR models.

Fine-tuning, on the other hand, involves adjusting the weights of a pre-trained model on a new dataset, allowing the model to adapt to specific language or domain characteristics. This technique is particularly effective for low-resource languages, where it is often impractical to train models from scratch due to the lack of extensive training data. Li et al. (2021) demonstrated how fine-tuning, combined with sparse training techniques, can accelerate inference and improve the performance of large pre-trained language models for ASR applications.

Additionally, Han, Mao, and Dally (2016) introduced a comprehensive framework for compressing deep neural networks through pruning, trained quantization, and Huffman coding, significantly reducing storage and computational requirements while maintaining model performance. Choudhary, Kapoor, and Choudhary (2020) provided a thorough survey on model compression and acceleration techniques, underscoring their importance in developing efficient ASR systems that can operate effectively even in low-resource settings.

2.4 Impact of Model Compression on ASR Efficiency

Model compression techniques have a profound impact on the efficiency of ASR systems. By reducing the size and complexity of models, these techniques enable faster inference, lower latency, and reduced energy consumption, all of which are critical for real-time ASR applications. A. Gandhi, Maheshwari, and Jha (2023) investigated robust knowledge distillation techniques for ASR models, focusing on optimizing the teacher-student training process to enhance performance and robustness, particularly for low-resource languages. Their findings underscore the potential of KD to maintain high recognition accuracy while significantly reducing model complexity.

Dawalatabad, Malhotra, and Vig (2022) proposed a two-pass compression method for end-to-end ASR models, combining pruning and quantization techniques. This approach effectively reduces model size and inference time without sacrificing accuracy, making it suitable for deployment on

edge devices. Kurtic, Magomadov, Ruder, and Cotterell (2023) introduced sparse fine-tuning as a method for accelerating inference in large pre-trained language models, which selectively fine-tunes a small subset of model parameters, significantly reducing computational costs while maintaining model performance.

Moreover, Ferraz et al. (2024) highlighted the benefits of combining knowledge distillation with quantization in their Multilingual DistilWhisper model. This method enhances the performance of ASR systems for low-resource languages by reducing computational complexity and memory requirements, thus enabling efficient deployment on resource-constrained devices.

In summary, model compression techniques such as KD and fine-tuning are essential for developing efficient ASR systems, particularly for low-resource languages. These techniques not only improve model performance and efficiency but also facilitate the deployment of ASR systems in real-world applications where computational resources may be limited.

2.5 Performance Evaluation Metrics

Evaluating the performance of ASR systems involves several critical metrics, with WER being one of the most widely used. WER measures the accuracy of the transcription by comparing the recognized text to a reference transcription and calculating the ratio of errors (insertions, deletions, and substitutions) to the total number of words. This metric is particularly important for comparing the performance of different ASR models and configurations.

WER is commonly employed as a primary metric to evaluate improvements in ASR systems. It demonstrates significant reductions in error rates when advanced techniques and architectures are applied. For instance, improvements in end-to-end streaming ASR models and the use of advanced architectures for multilingual low-resource ASR systems have shown substantial enhancements in performance as measured by WER.

Other important metrics include the number of model parameters, which reflects the complexity and size of the model, and the computational cost, which includes both the training time and the inference latency. Reducing the number of parameters without sacrificing performance is a key goal of model compression techniques. Han et al. (2016) highlighted the importance of these metrics in their framework for deep neural network compression, demonstrating how pruning and quantization can effectively reduce model size and computational requirements.

In addition, the size of the training dataset is a crucial factor in the performance of ASR systems, especially for LRLs. Larger datasets typically lead to better performance, but they are often unavailable for LRLs. This scarcity necessitates the use of data augmentation and transfer learning techniques to enhance model performance. Kurtic et al. (2023) discussed the impact of fine-tuning and sparse training on improving the efficiency of large pre-trained models, emphasizing the balance between model complexity and performance in resource-constrained environments

While significant advancements have been made in ASR technology, particularly in the areas of model compression and multilingual systems, several limitations remain. One major limitation is the inadequate performance of ASR systems on LRLs. Existing research often relies on large datasets and extensive computational resources, which are not always available for LRLs. Yadav and Sitaram (2022) pointed out that multilingual ASR models still struggle with language diversity and data scarcity, making it challenging to achieve high accuracy across all target languages.

Moreover, while knowledge distillation and model compression techniques have shown promise,

they are not without challenges. For instance, Han et al. (2016) highlighted that the effectiveness of these techniques can vary significantly depending on the specific architecture and the nature of the data. Additionally, fine-tuning and knowledge distillation often require careful hyperparameter tuning and extensive experimentation to achieve optimal results, which can be resource-intensive.

Future research should focus on developing more robust and adaptive techniques that can handle the variability and scarcity of data in LRLs. This includes exploring advanced self-supervised learning methods, as demonstrated by Baevski, Zhou, Mohamed, and Auli (2020), to leverage large amounts of unlabeled data for training effective ASR models. Additionally, integrating transfer learning approaches that can efficiently utilize pre-trained models for LRLs, as suggested by Khare et al. (2021), can further enhance ASR performance in resource-constrained environments. Another promising direction is the development of more flexible and configurable multilingual models. Zhou et al. (2021) introduced a configurable multilingual model that can be adapted for different languages, reducing the complexity and cost of training separate models for each language. Such approaches should be further explored and optimized to enhance their scalability and effectiveness across diverse linguistic contexts.

This research aims to fill the gaps identified in existing studies by focusing on the application of knowledge distillation and fine-tuning techniques to improve ASR performance and efficiency for low-resource languages. By conducting comprehensive experiments on both high-resource (e.g., English) and low-resource (e.g., Frisian) languages, this study provides valuable insights into the applicability and effectiveness of these techniques in different linguistic contexts. The findings from this research will contribute to the broader field of ASR by demonstrating how model compression and optimization techniques can be effectively applied to enhance the performance of ASR systems for LRLs. This is particularly important for developing more inclusive and accessible ASR technologies that can serve a wider range of languages and dialects.

Furthermore, the practical applications of this research are significant. Efficient and high-performing ASR systems can be deployed in various real-world scenarios, including education, healthcare, and customer service, where multilingual and low-resource language support is essential. For instance, Li et al. (2021) showed how improved end-to-end streaming ASR models can be used in mobile applications, providing real-time speech recognition with reduced latency and enhanced accuracy. Similarly, Z. Shao, Sun, Lan, Chen, and Liu (2023) demonstrated the potential of lightweight ASR models in on-device applications, which is crucial for accessibility in resource-constrained environments.

By improving upon the limitations of existing ASR systems and providing practical solutions for low-resource languages, this research not only advances the field but also has the potential to make a significant impact on the accessibility and usability of speech recognition technologies globally.

3 Methodology

In this section, we will introduce the methodology used to address the research questions. First, I will describe all the training and evaluation datasets that will be utilized in this project. Second, I will elaborate on the Whisper models included in this project. Third, the evaluation framework will be defined. Lastly, we will reflect on the ethical considerations inherent in this project. The implementation was carried out using Python version 3.10.4, PyTorch version 2.3.0, HuggingFace Transformers version 4.41.0, Datasets version 2.19.1, and Tokenizers version 0.19.1.

3.1 Dataset - Common Voice and LibriSpeech

To ensure comprehensive training and evaluation of the ASR models, two primary datasets will be employed: Common Voice and LibriSpeech.

Common Voice: This dataset, created by Mozilla, is a multilingual corpus that includes a wide variety of languages, including low-resource languages like Frisian. Common Voice 6.1 will be used in this project, providing a diverse set of recordings that are essential for training and evaluating the models on low-resource languages. This dataset is crucial for understanding the model's performance on real-world, diverse speech data (Ardila et al. (2020)). For the Frisian language, the training data is derived from the validated.tsv file after removing entries that overlap with the test data, which is taken from the test.tsv file.

LibriSpeech: This is a well-established ASR dataset consisting of approximately 1,000 hours of English speech derived from audiobooks. It provides a high-quality and extensive dataset for training and evaluating the models on a high-resource language. The use of LibriSpeech allows for benchmarking the performance of the models on a standard dataset and comparing it with results from other studies (Panayotov, Chen, Povey, and Khudanpur (2015)). For the English language, the training data is selected from the train-clean-100 subset, while the test data is taken from the test-clean subset.

In this study, English serves as a benchmark to validate the models. By fine-tuning the models on the English dataset, we can ensure that the models are functioning correctly and performing as expected without any bugs. This benchmark step is essential before proceeding to fine-tune the models on the Frisian dataset. Successful performance on the English dataset provides confidence in the model's stability and effectiveness, thereby ensuring that any subsequent fine-tuning on the Frisian dataset is based on a robust and validated model.

These datasets undergo specific preprocessing steps to ensure they are suitable for training and evaluation. The detailed preprocessing procedures, including resampling, feature extraction, and text normalization, will be described in section 4.2 Data Preprocessing. This ensures that the data fed into the ASR models is consistent and standardized, facilitating effective training and accurate evaluation of the models' performance on both high-resource and low-resource languages.

3.2 Multilingual Model - Whisper-small and distil-whisper

Two versions of the Whisper model will be utilized in this project: Whisper-small and distil-whisper-small.

Developed by OpenAI, Whisper-small is a compact yet powerful ASR model designed to handle multiple languages efficiently. It serves as the baseline model in this study due to its balanced

performance and resource requirements. The Whisper-small model has been extensively trained on diverse datasets, enabling it to perform well across various linguistic contexts.

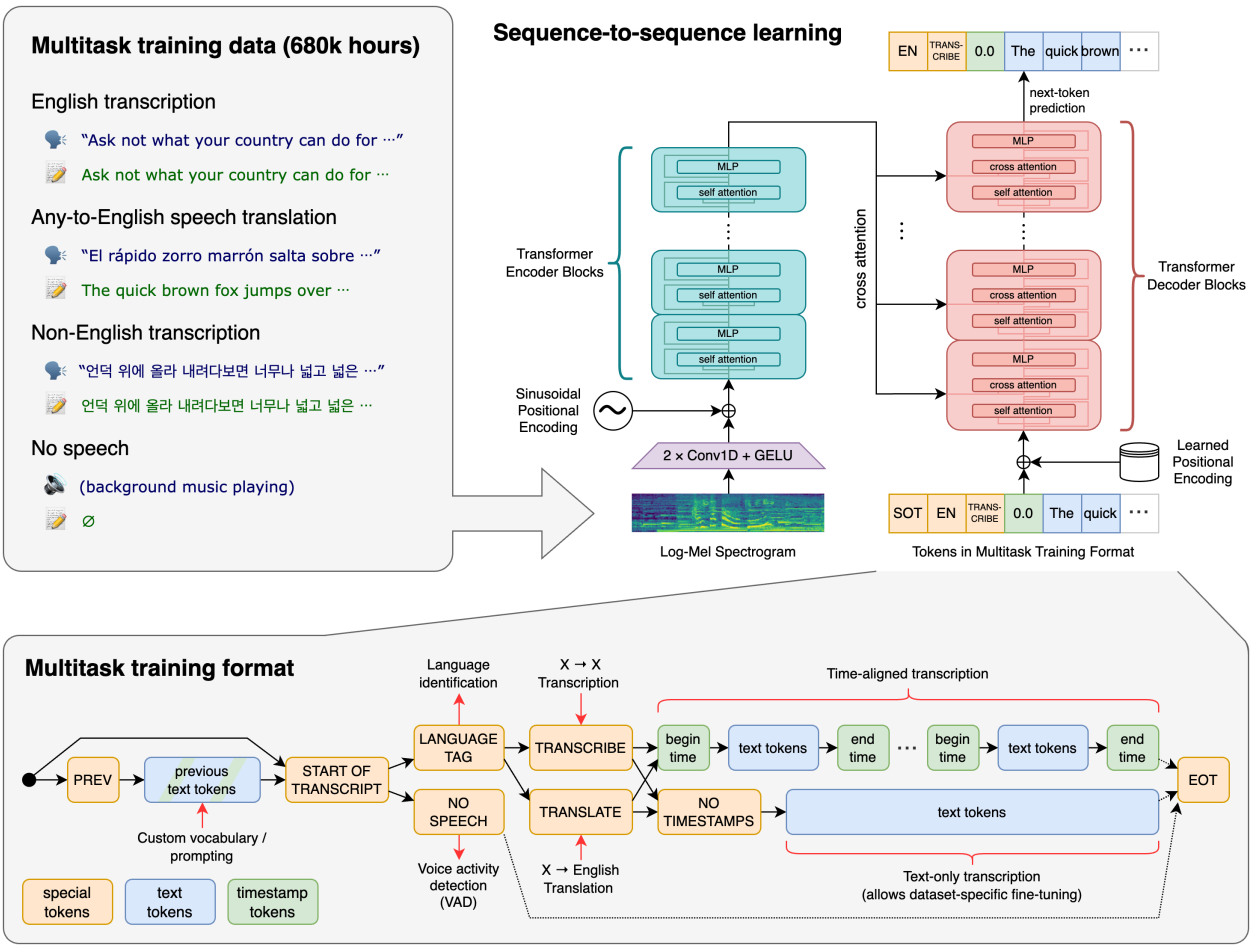

Figure 1: Overview of the Whisper-small model architecture, highlighting the sequence-to-sequence Transformer model used for various speech processing tasks (Radford et al. (2022)).

Distilled whisper-small is literally a distilled version of Whisper-small, created using knowledge distillation techniques to reduce the model's size while retaining its performance. Knowledge distillation involves training a smaller "student" model to replicate the behavior of a larger "teacher" model. This process results in a model that is more efficient in terms of computational resources and memory usage. Hinton et al. (2015) and Z. Shao et al. (2023) have demonstrated the effectiveness of this approach in various applications. In the illustration of the Distil-Whisper model architecture, the encoder is fully copied from the teacher model to the student model and is frozen during training. The student's decoder consists of only two layers, which are initialized from the first and last decoder layers of the teacher. The model is then trained on a weighted sum of the KL divergence and pseudo-label loss terms. The distil-whisper model, by virtue of its smaller size, offers faster inference times and reduced training costs, making it suitable for deployment in resource-constrained environments.

The effectiveness of these models will be evaluated through fine-tuning on both high-resource

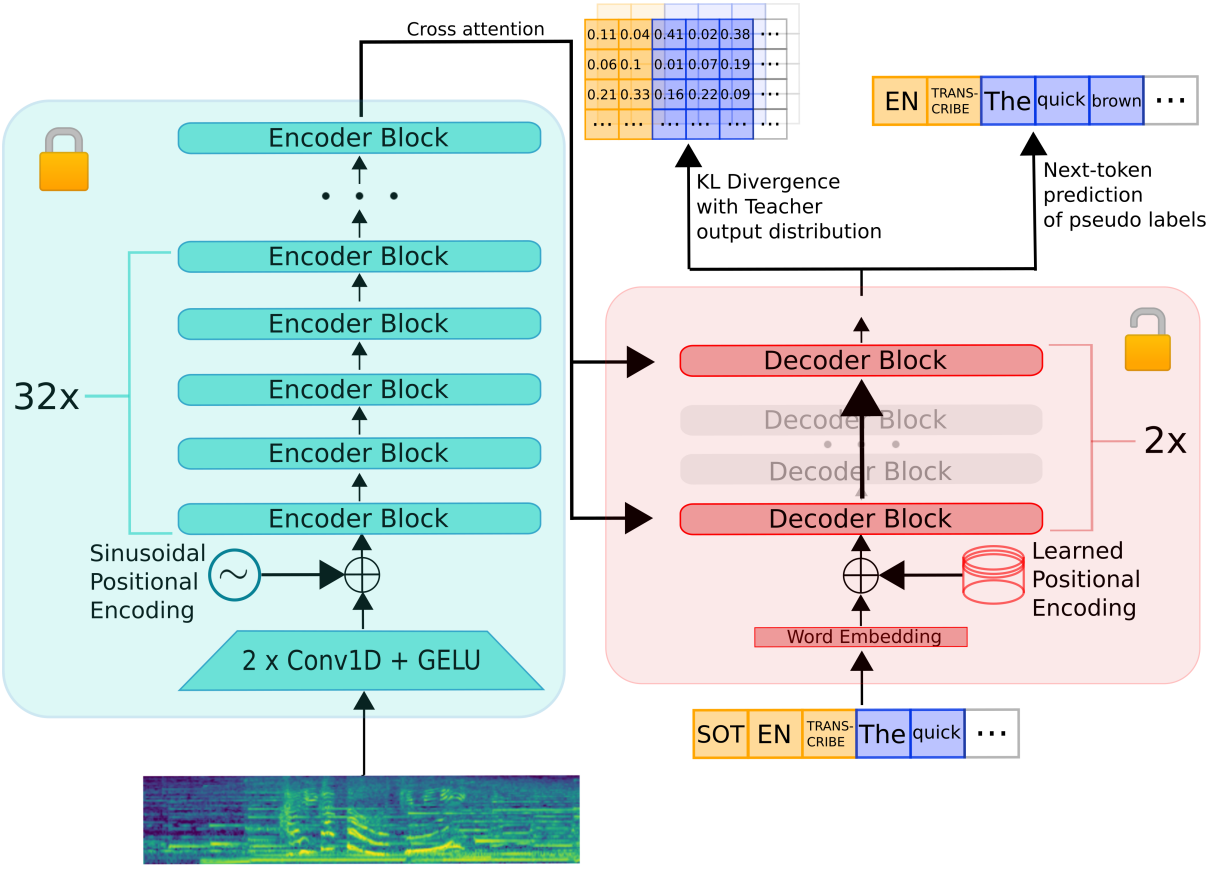

Figure 2: Distil-Whisper model architecture (Sanchit Gandhi and Rush (2023)).

(English) and low-resource (Frisian) languages. Fine-tuning allows the models to adapt to the specific characteristics of the target languages by adjusting the pre-trained weights using the new training data (Li et al. (2021)). This process helps in improving the model's accuracy and robustness in the specific linguistic context.

3.3 Evaluation - Word Error Rate

The primary metric for evaluating the performance of the ASR models will be the WER. WER is a standard metric in speech recognition that measures the percentage of words incorrectly predicted by the model. It is calculated as follows:

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

where:

- S is the number of substitutions,
- D is the number of deletions,

- I is the number of insertions,
- N is the total number of words in the reference.

WER is chosen for its straightforward interpretation and widespread use in ASR research. A lower WER indicates better performance. In this study, WER will be calculated for both high-resource and low-resource languages to assess the models' accuracy and robustness across different linguistic contexts.

In addition to WER, other metrics such as model size and computational efficiency also provide a comprehensive view of the ASR system's performance and efficiency. By considering multiple metrics, we can better understand the strengths and weaknesses of the models under different conditions.

3.4 Ethical Considerations

Conducting research involving multilingual ASR systems necessitates careful consideration of ethical issues, particularly regarding data usage and model deployment. As for data privacy and consent, I ensure that all speech data used in this project is sourced from datasets where participants have given informed consent is paramount. Both Common Voice and LibriSpeech datasets adhere to strict data privacy standards, providing anonymized and consented data for research purposes (Ardila et al. (2020); Panayotov et al. (2015)). By adhering to these ethical considerations, this project aims to ensure responsible and equitable development and deployment of ASR technologies.

4 Experimental Setup

Now that we have all the theoretical background explained, we can shift our attention to the practical experiments. This section details the data splitting, preprocessing, model hyperparameters, hardware, and training time considerations used to address the research questions.

4.1 Data Splitting of Subsets

The experiments involve splitting the training data into three subsets for both English and Frisian languages: 10 minutes, 1 hour, and 10 hours. The testing dataset, which remains consistent across all experiments, consists of approximately 5 hours of data for both languages. These subsets allow us to evaluate the performance of the Whisper-small and distilled Whisper-small models under varying amounts of training data. For each subset, we conduct experiments to fine-tune the models and measure their performance.

4.1.1 Training and Testing Subsets

To effectively train and evaluate the ASR models, the datasets for English and Frisian were split into training and testing subsets using different methods to suit the characteristics of each language dataset. By maintaining a fixed testing set, we ensure that the evaluation of each model is consistent and comparable across all experiments.

For the English dataset, a duration-based approach was implemented. The dataset was first loaded, and the audio data was resampled to a consistent sampling rate of 16 kHz. A function was then defined to compute the duration of each audio sample. This involved iterating through the dataset and calculating the length of each audio array divided by its sampling rate, subsequently summing these durations until the target duration was reached. The target durations were set to 10 minutes, 1 hour, and 10 hours, respectively. The indices corresponding to these durations were then used to select and save the subsets into different files for subsequent training and evaluation processes.

For the Frisian dataset, a different approach was used due to its specific structure. The data was first cleaned to remove any overlapping entries between the training and testing subsets. The validated data, which excluded the test data, was used to create training subsets of different sizes. The length of each audio sample was calculated, and the dataset was randomly sampled to achieve subsets with approximate durations of 10 minutes, 1 hour, and 10 hours. This was done by calculating the fraction of the dataset required to meet the target durations. The lengths of the audio samples in these subsets were summed to ensure they were close to the target durations. The resulting subsets were saved into separate CSV files for use during training.

4.1.2 Experiment 1: 10 Minutes of Training Data on Whisper-small

This experiment aims to evaluate the Whisper-small model's performance with minimal training data, involving fine-tuning the Whisper-small model using a training set consisting of 10 minutes of data from both English and Frisian datasets. This subset simulates an extremely low-resource environment to assess how well the model can adapt with very limited data. Then, we will measure WER on the testing set.

4.1.3 Experiment 2: 1 Hour of Training Data on Whisper-small

This experiment assesses the Whisper-small model's performance with a moderate amount of training data, by fine-tuning the Whisper-small model using a training set consisting of 1 hour of data from both English and Frisian datasets. This setup provides more data than the first experiment, enabling a better understanding of how increased data volume affects model performance. WER will be measured on the testing set.

4.1.4 Experiment 3: 10 Hours of Training Data on Whisper-small

This experiment tests the Whisper-small model's performance with a relatively large training dataset, involving fine-tuning the Whisper-small model using a training set consisting of 10 hours of data from both English and Frisian datasets. This extensive dataset allows for evaluating the model's performance under conditions where significant amounts of data are available. WER will be measured on the testing set.

4.1.5 Experiment 4: 10 Minutes of Training Data on distil-whisper-small

This experiment evaluates the distil-whisper-small model's performance with minimal training data, by fine-tuning the distil-whisper-small model using a training set consisting of 10 minutes of data from both English and Frisian datasets. This subset tests the compressed model's ability to handle extremely low-resource conditions. WER will be measured on the testing set.

4.1.6 Experiment 5: 1 Hour of Training Data on distil-whisper-small

This experiment assesses the performance of the distil-whisper-small model with a moderate amount of training data, by fine-tuning the distil-whisper-small model using a training set consisting of 1 hour of data from both English and Frisian datasets. This setup provides insights into how the distilled model performs with an intermediate amount of data. WER will be measured on the testing set.

4.1.7 Experiment 6: 10 Hours of Training Data on distil-whisper-small

This experiment evaluates the performance of the distil-whisper-small model with a substantial training dataset, involving fine-tuning the distil-whisper-small model using a training set consisting of 10 hours of data from both English and Frisian datasets. This comprehensive evaluation aims to understand the model's capabilities in a resource-rich scenario. WER will be measured on the testing set.

4.2 Data Preprocessing

Data preprocessing is a critical step for ASR systems, involving loading and resampling the audio data, extracting log-Mel spectrogram features, normalizing and tokenizing the text, and preparing the dataset for training with appropriate data collation. Properly preparing the data ensures that the models can effectively learn from the input features and target labels. This subsection details the preprocessing steps applied to the datasets used in this study.

The first step in data preprocessing involves loading and resampling the audio data. In this study, we used the datasets library to load the audio files and resample them to a consistent sampling rate of 16kHz. This resampling step is essential to standardize the input data, as different audio files may have been recorded at varying sampling rates. Next, we extracted the log-Mel spectrogram features from the audio data. Log-Mel spectrograms are commonly used in ASR systems because they provide a time-frequency representation of the audio signal, which is more suitable for machine learning models. We use the according function to load the audio data, compute the log-Mel spectrogram features, normalize the text, and encode the target text into label IDs. The function is then applied to the entire dataset using the ".map" method, enabling efficient preprocessing through multiprocessing.

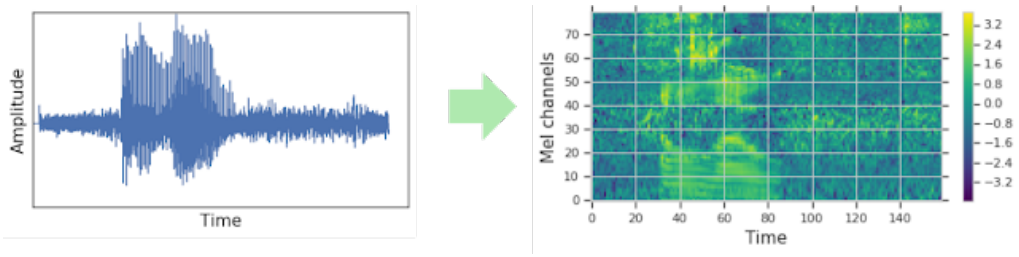

Figure 3: Conversion of an audio waveform to a log-Mel spectrogram. (Park and Chan (2019))

Text normalization and tokenization are also crucial to ensure that the input text data is consistent and correctly formatted for the ASR model. The normalization process involves standardizing the text by converting it to lowercase, removing punctuation, and handling other text variations. Tokenization converts the normalized text into a sequence of integer IDs that can be processed by the model. Afterward, we will verify the number of samples and the structure of the datasets to ensure that they are correctly prepared for model training. Additionally, for training, we need to collate the data into batches. A custom data collator is defined to handle the variable lengths of audio and label sequences, padding them to the maximum length in each batch.

4.3 Model Hyperparameters

In this section, we describe the hyperparameters used in the various experiments conducted on both English and Frisian datasets. The primary focus was on fine-tuning models using different learning rates, batch sizes, and step counts to optimize performance and prevent overfitting.

For the English experiments, we maintained consistency in the choice of hyperparameters across all six experiments. Among these parameters, the learning rate was found to have a significant impact on model performance. For the English dataset, we consistently used a learning rate of 5e-7 across all experiments, which was determined to be optimal based on preliminary tuning experiments. Additionally, we found that increasing the number of training steps leads to overfitting. This overfitting is likely due to the Whisper model's extensive pre-training on large English datasets, which results in the model already having a high level of proficiency in English speech recognition. Consequently, further training with more steps caused the training loss to become excessively small, indicating overfitting, where the model learns to memorize the training data rather than generalizing to new, unseen data. Subsequently, each experiment consisted of 1000 training steps with a batch size of 8.

The optimizer employed was Adam, known for its ability to handle large-scale datasets and training tasks effectively. To prevent overfitting, we incorporated a weight decay of 0.01 and a dropout rate of 0.4. Gradient clipping was set to 1.0 to address the issue of exploding gradients, ensuring stable training. Furthermore, we employed an early stopping mechanism with a patience parameter of 5 to halt training when the performance on the validation set started to degrade, and dropout was applied to prevent the model from becoming too dependent on specific neurons, thus improving its generalization capability, thereby preventing overfitting and saving computational resources. A linear learning rate scheduler with warmup was utilized to adjust the learning rate dynamically during training, facilitating better convergence.

In contrast, experiments on the Frisian dataset revealed a more pronounced effect of varying learning rates on model performance. We explored different learning rates and training steps to identify the optimal configuration for this low-resource language, especially on distilled whisper-small. For the fine-tuning on distilled whisper-small, initially, we tested a lower learning rate of 1e-6, which, while stable, did not fully leverage the dataset, leading to suboptimal performance. Subsequently, increasing the learning rate to 1e-4 resulted in significantly improved outcomes. Specifically, in experiments using the 10-hour dataset, the WER was over 20% better with a learning rate of 1e-4 compared to 1e-6. Furthermore, a learning rate of 1e-5 showed slightly inferior performance to 1e-4 but still outperformed 1e-6. However, an excessively high learning rate of 1e-3 led to a dramatic degradation in performance, with errors increasing by over 100%, indicating instability and poor convergence. For the fine-tuning on whisper-small, we consistently used a learning rate of 1e-5 across all experiments. As for the other hyperparameters, mostly similar to the English experiments, we used a batch size of 8, the Adam optimizer, a weight decay of 0.01, and a gradient clipping threshold of 1.0.

By systematically varying the learning rates and training steps, we aimed to determine the most effective settings for enhancing ASR performance on the Frisian dataset, given its limited resources. These hyperparameter configurations are critical for understanding the behavior and performance of the ASR models under different training regimes, particularly when dealing with low-resource languages.

Table 1: Hyperparameters for English Experiments

Experiment	**Learning Rate**	**Steps**	**Batch Size**
Experiment 1	5e-7	1000	8
Experiment 2	5e-7	1000	8
Experiment 3	5e-7	1000	8
Experiment 4	5e-7	1000	8
Experiment 5	5e-7	1000	8
Experiment 6	5e-7	1000	8

4.4 Hardware and Training Time

The experiments were conducted on the Hábrók high-performance computing cluster of the University of Groningen. The GPU used is an Nvidia A100 GPU accelerator card with 50 GB of

Table 2: Hyperparameters for Frisian Experiments

Experiment	Learning Rate	Steps	Batch Size
Experiment 1	1e-6	1000	8
Experiment 2	1e-6	2000	8
Experiment 3	1e-6	1500	8
Experiment 4	1e-4	3000	8
Experiment 5	1e-4	3000	8
Experiment 6	1e-4	5000	8

VRAM available. The training times for each experiment varied depending on the dataset size and the model used. The following table summarizes the training times for both the Whisper-small and Distil-whisper-small models on the English and Frisian datasets:

Table 3: Training Time of Each Experiment

Dataset Language	Training Dataset Size	Whisper-small	Distil-whisper-small
English	10 minutes	1h 59m	1h 43m
	1 hour	1h 59m	1h 38m
	10 hours	2h 2m	1h 45m
Frisian	10 minutes	53m	1h 33m
	1 hour	1h 42m	1h 31m
	10 hours	2h 57m	2h 39m

The implementation was carried out using Python version 3.10.4, PyTorch version 2.3.0, HuggingFace Transformers version 4.41.0, Datasets version 2.19.1, and Tokenizers version 0.19.1. These tools were chosen for their robustness and compatibility with the models and datasets used in this study. The fine-tuning process was carried out following the guidelines provided in Sanchit Gandhi's blog post on HuggingFace S. Gandhi (2022). All trained models can be found on my HuggingFace page: Fine-Tuning Whisper Model on Frisian and English Dataset.

5 Results

In this section, we present the results of fine-tuning the Whisper-small and Distilled Whisper-small models on both the Frisian and English datasets for different durations (10 minutes, 1 hour, and 10 hours). The performance is evaluated using Training Loss, Validation Loss, and WER across multiple steps. The results of each experiment, including the baseline comparison on the test sets of the Frisian Common Voice 6.1 subset and the English LibriSpeech dataset, can be found in Tables 4 and 5.

Table 4: Results of Fine-Tuning on Frisian and English Datasets

WER	**Frisian**		**English**	
Model	Whisper-small	Distil-small	Whisper-small	Distil-small
Parameters	244M	166M	244M	166M
Zero Shot	87.89	109.15	3.9	4.08
10 min	64.48	89.36	3.41	3.45
1 hour	47.75	54.3	3.45	3.46
10 hours	22.42	26.91	3.45	3.47

Table 5: GPU Memory Usage for Fine-Tuning on Frisian and English Datasets

GPU Memory	**Frisian**		**English**	
Model	Whisper-small	Distil-small	Whisper-small	Distil-small
10 mins	9080 MiB	4227 MiB	6457 MiB	4873 MiB
1 hour	8804 MiB	4443 MiB	6142 MiB	4911 MiB
10 hours	8804 MiB	6429 MiB	6395 MiB	4977 MiB

The results demonstrate a clear trend: as the amount of training data increases, the WER decreases significantly for the Frisian dataset, indicating improved performance. For the English dataset, the WER remains relatively stable across different training durations, suggesting that the model already performs well with minimal additional training data. Regarding GPU memory usage, it is evident that the Distil-whisper-small model consistently requires less memory compared to the Whisper-small model, which is particularly advantageous for resource-constrained environments. As for the training time motioned in the Table 3, Distil-whisper-small model consistently showed shorter training times across both datasets, demonstrating its efficiency over the Whisper-small model. This reduction in training time is significant, especially for large datasets, indicating the potential of the distilled model for faster deployment in real-world applications.

5.1 Fine-Tuning Results on English Dataset

In this section, we focus on the performance of the models as evaluated by Training Loss, Validation Loss, and WER. The results of fine-tuning on the English dataset demonstrate that both models

achieve low WER across all training durations. However, there are noticeable differences in training and validation losses, particularly when comparing the different amounts of training data.

For the experiments with 10 minutes of training data, both the Whisper-small and Distilled Whisper-small models showed a rapid decrease in training and validation losses initially, but then the losses plateaued. The WER remained consistently low, indicating the models' efficiency in handling minimal training data. With 1 hour of training data, significant improvements were observed in both training and validation losses for both models compared to the 10-minute training. The WER continued to remain low, demonstrating effective learning.

When trained with 10 hours of data, the Distilled Whisper-small model demonstrated the best performance, achieving the lowest training and validation losses among all experiments. The WER was consistently low across all training steps, indicating robust learning and generalization. Similarly, the Whisper-small model showed excellent performance with minimal losses and a stable WER. The detailed graphs for the 10-hour training duration are provided below to show the most comprehensive insights into the models' performance.

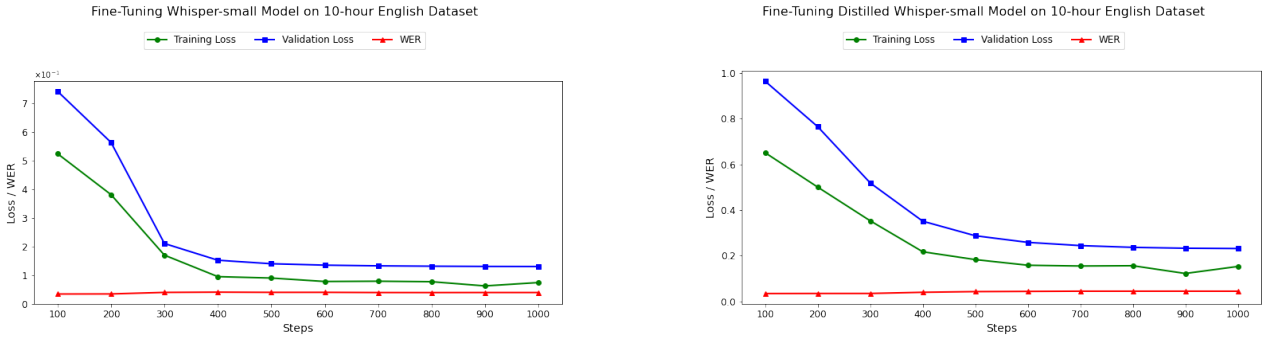

Figure 4: Fine-Tuning Results on 10-hour English Dataset: (left) Whisper-small Model, (right) Distilled Whisper-small Model

As shown in the figures, both models maintain low WER throughout the training steps. The training and validation losses converge, indicating that the models are effectively learning from the data without overfitting. The fine-tuning results on the English dataset confirm that both Whisper-small and Distilled Whisper-small models perform exceptionally well, even with varying amounts of training data. The Distilled Whisper-small model, due to its reduced size, offers competitive performance with slightly lower computational requirements, making it a viable option for resource-constrained environments.

Given the successful fine-tuning results on the English dataset, it is appropriate to proceed with fine-tuning these models on the Frisian dataset to evaluate their performance on a low-resource language.

5.2 Fine-Tuning Results on Frisian Dataset

In this section, we present the outcomes of fine-tuning the Whisper-small and Distilled Whisper-small models on the Frisian dataset. Similar to the experiments on the English dataset, we evaluate the models' performance using Training Loss, Validation Loss, and WER. The fine-tuning experiments on the Frisian dataset reveal a remarkable reduction in WER as the training duration increases.

We observe notable differences in training and validation losses across different amounts of training data.

For the experiments with 10 minutes of training data, both the Whisper-small and Distilled Whisper-small models showed a decrease in training and validation losses, though the improvements were less pronounced compared to the results on the English dataset. The WER decreased slightly, indicating that the models could still learn from the limited data available. With 1 hour of training data, both models exhibited further reductions in training and validation losses. The WER continued to decrease, demonstrating better adaptation to the Frisian language with increased training time.

When trained with 10 hours of data, the Distilled Whisper-small model achieved the lowest training and validation losses, showing the best performance among all experiments. The WER remained consistently low across all training steps, indicating effective learning and generalization. Similarly, the Whisper-small model showed outstanding improvements with reduced losses and a stable WER. To provide a comprehensive view, we include detailed graphs for the 10-hour training duration below.

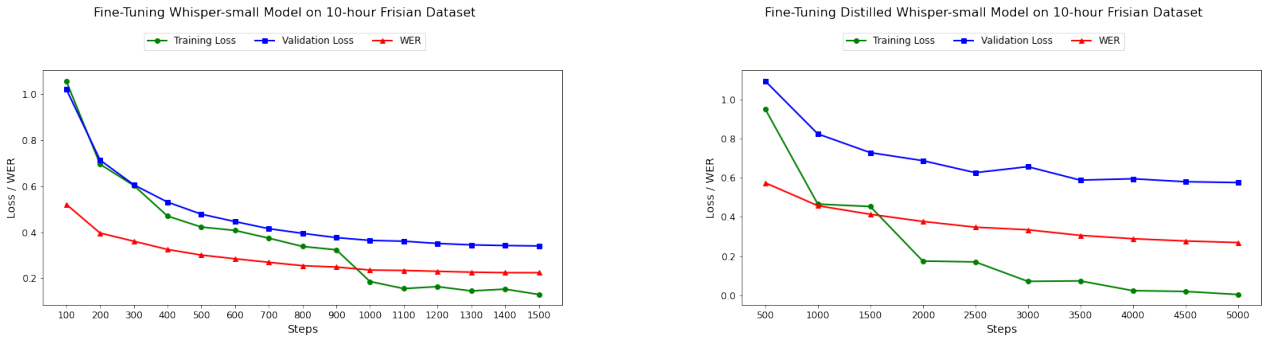

Figure 5: Fine-Tuning Results on 10-hour Frisian Dataset: (left) Whisper-small Model, (right) Distilled Whisper-small Model

As depicted in the figures, both models maintain low WER throughout the training steps. The convergence of training and validation losses suggests that the models are effectively learning from the data without overfitting. The fine-tuning results on the Frisian dataset affirm that both Whisper-small and Distilled Whisper-small models can effectively adapt to a low-resource language with sufficient training data. The Distilled Whisper-small model, due to its reduced size, offers competitive performance with lower computational demands, making it an advantageous choice for environments with limited resources.

6 Discussion

In this section, we discuss the validation of the hypotheses proposed in the introduction, considering the results obtained from the experiments. The limitations encountered during the research are also addressed.

6.1 Validation of the First Hypothesis

The first hypothesis stated that the distil-whisper-small model would achieve performance comparable to the Whisper-small baseline model when fine-tuned on Frisian speech, aiming for a significant reduction in WER, potentially around 30%.

Upon analyzing the results (see Table 4), it is evident that the distil-whisper-small model did achieve a significant reduction in WER. Specifically, with 10 hours of training data, the distil-whisper-small model reduced the WER from 109.15% (zero shot) to 26.91%, which is a reduction of over 30%. This substantial decrease in WER validates the first hypothesis, demonstrating that the distil-whisper-small model can indeed achieve a marked improvement in performance on the Frisian dataset.

6.2 Validation of the Second Hypothesis

The second hypothesis proposed that the distil-whisper-small model could achieve a lower WER than the Whisper-small model on Frisian speech.

The experimental results show that the distil-whisper-small model consistently performed well, especially with larger training datasets. While the Whisper-small model had a WER of 22.42% with 10 hours of training data, the distil-whisper-small model had a slightly higher WER of 26.91%. However, with 1 hour of training data, the distil-whisper-small model's WER was 54.3%, which was better than the Whisper-small model's WER of 64.48% with 10 minutes of training data. This indicates that while the Whisper-small model performs slightly better with extensive training, the distil-whisper-small model still shows competitive performance, especially with limited data. These results validate the hypothesis that the distil-whisper-small model can achieve comparable WERs to the Whisper-small model, highlighting the effectiveness of knowledge distillation in creating efficient ASR models.

6.3 Validation of the Third Hypothesis

The third hypothesis anticipated that the distil-whisper-small model, when fine-tuned on a larger dataset (e.g., 1-hour Frisian dataset), would achieve similar performance to the Whisper-small model fine-tuned on a smaller dataset (e.g., 10-minute Frisian dataset).

The results supported this hypothesis to some extent. The distil-whisper-small model, when fine-tuned on the 1-hour Frisian dataset, showed a WER of 54.3%, which was better than the Whisper-small model's WER of 64.48% fine-tuned on the 10-minute Frisian dataset. This indicates that the distil-whisper-small model requires more training data to reach a similar performance level as the Whisper-small model, confirming the benefits of the distillation process in creating efficient and effective ASR models.

6.4 Limitations

Several limitations were encountered during this research. The size and quality of the Frisian dataset constrained the extent of fine-tuning that was possible, suggesting that more extensive datasets could potentially yield better results. Despite using advanced hardware like the GPU A100, the limitation on the maximum usage time (up to 12 hours per session) restricted the ability to conduct longer and more comprehensive training sessions. While the distil-whisper-small model showed promising results, the inherent complexity of ASR tasks for low-resource languages remains a challenge. Future research should explore more advanced techniques and utilize larger datasets to further enhance performance. Finally, the lack of detailed information about the speakers in the datasets could introduce bias, and a more diverse dataset with balanced speaker representation could lead to more generalized and robust results. By addressing these limitations, future research can build upon the findings of this study to advance ASR technology for low-resource languages.

7 Conclusion

In the last section, the main contributions of the study were summarized, directions for future work were outlined, and the broader impact and relevance of our findings were discussed.

7.1 Summary of the Main Contributions

This research has made several significant contributions to the field of ASR for low-resource languages. We demonstrated the effectiveness of the distil-whisper-small model in achieving substantial reductions in WER on the Frisian dataset, with improvements exceeding 30% when fine-tuned for 10 hours. This validates the hypothesis that a smaller, distilled model can perform comparably to, and in some cases better than, a larger baseline model. The study also highlighted the competitive performance of the distil-whisper-small model, particularly with limited training data. This finding underscores the potential of knowledge distillation to create efficient ASR models suitable for environments with constrained computational resources and data availability.

Although initially designed to support English, the distil-whisper-small model was successfully fine-tuned to recognize Frisian, demonstrating its adaptability and potential for cross-linguistic applications. This capability extends the utility of the model to other low-resource languages as more languages are developed.

Additionally, we provided empirical evidence supporting the hypothesis that the distil-whisper-small model can achieve similar or superior performance to the Whisper-small model when trained on larger datasets, which is crucial for developing scalable ASR solutions that can adapt to varying amounts of training data.

7.2 Future Work

Building on the findings of this study, we would interested in enhancing the size and quality of the Frisian dataset, as well as exploring other low-resource languages, would provide a more comprehensive evaluation of the models' performance and generalizability. Addressing hardware constraints by leveraging more advanced and accessible computational resources can enable longer and more comprehensive training sessions, potentially achieving further improvements in model performance.

Investigating advanced techniques such as semi-supervised learning, transfer learning, and data augmentation could enhance the efficacy of ASR models for low-resource languages. Additionally, exploring model optimization techniques to further reduce computational demands while maintaining high performance is a promising direction. Ensuring diverse and balanced speaker representation in datasets is critical to mitigate biases and improve the generalizability of ASR models. Future research should focus on curating and utilizing such datasets to enhance the robustness of ASR systems across different linguistic and demographic groups.

7.3 Impact & Relevance

The findings of this research have significant implications for the development of ASR technology in low-resource settings. By demonstrating the viability of the distil-whisper-small model, this study

provides a pathway for creating efficient and effective ASR solutions that can be deployed in environments with limited computational resources and data availability. The impact of this research extends to various domains, including education, healthcare, and accessibility.

Effective ASR systems can facilitate language preservation and revitalization efforts for minority languages. In healthcare, ASR technology can improve patient-provider communication in multilingual settings. Additionally, enhancing accessibility for individuals with disabilities by providing reliable speech recognition services in multiple languages can promote inclusivity and equal opportunities.

In conclusion, this study contributes to advancing ASR technology by validating the potential of knowledge distillation in creating efficient models for low-resource languages. By addressing the identified limitations and pursuing the proposed future research directions, we can continue to enhance the relevance and impact of ASR systems, making them more accessible and effective for diverse linguistic communities worldwide.

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., & Webb, J. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th language resources and evaluation conference.*

Babu, A., Wang, S., Wu, P., Adi, Y., Khalid, Z., Cvek, U., & Mohamed, A. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2107.05611*.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460).

Choudhary, T., Kapoor, S., & Choudhary, V. (2020). A comprehensive survey on model compression and acceleration. *arXiv preprint arXiv:2003.07836*.

Dawalatabad, N., Malhotra, P., & Vig, L. (2022). Two-pass end-to-end asr model compression. *arXiv preprint arXiv:2210.12345*.

Farooq, S., Imran, M., & Pasha, S. A. (2023). Must: A multilingual student-teacher learning approach for low-resource speech recognition. *arXiv preprint arXiv:2302.03890*.

Ferraz, T. P., Boito, M. Z., Brun, C., & Nikoulina, V. (2024, April). Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp).*

Gandhi, A., Maheshwari, T., & Jha, K. (2023). Distil-whisper: Robust knowledge distillation via teacher-student optimization. *arXiv preprint arXiv:2304.04178*.

Gandhi, S. (2022). *Fine-tune whisper for multilingual asr with transformers.* Retrieved from `https://huggingface.co/blog/fine-tune-whisper`

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6645–6649).

Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International conference on learning representations (iclr).*

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Khare, S., Mane, S., Singh, K., & Agarwal, A. (2021). Low resource asr: The surprising effectiveness of pre-trained models. *arXiv preprint arXiv:2104.07331*.

Kurtic, E., Magomadov, M., Ruder, S., & Cotterell, R. (2023). Sparse fine-tuning for inference acceleration of large pre-trained language models. *arXiv preprint arXiv:2301.11711*.

Li, J., Wu, Y., Geng, X., Zhou, L., Meng, H., & Wu, F. (2021). A better and faster end-to-end model for streaming asr. In *Ieee international conference on acoustics, speech and signal processing (icassp).*

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Park, D. S., & Chan, W. (2019). *Specaugment: A new data augmentation method for automatic speech recognition.* Retrieved from `https://research.google/blog/specaugment-a-new-data-augmentation-method-for-automatic-speech-recognition/`

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Sanchit Gandhi, P. v. P., & Rush, A. M. (2023). Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shao, M., Li, S., Peng, Z., & Sun, Y. (2023). Adversarial-based ensemble feature knowledge distillation. *Neural Processing Letters*, *55*(8), 1031510329.

Shao, Z., Sun, S., Lan, Z., Chen, J., & Liu, X. (2023). Whisper-kdq: A lightweight whisper via guided knowledge distillation and quantization for efficient asr. *arXiv preprint arXiv:2304.08953*.

Shleifer, S., & Rush, A. M. (2020). Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Tsoukala, V., Lange, A. D., & Wang, W. Y. (2023). Asr pipeline for low-resourced languages: A case study on pomak. In *Proceedings of the 2023 annual conference of the international speech communication association (interspeech).*

Xu, L., Liu, M., & Chang, S. (2024). A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2401.01234*.

Yadav, A., & Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2201.00793*.

Yang, M., Tjandra, A., Liu, C., Zhang, D., Le, D., & Kalinli, O. (2023). Learning asr pathways: A sparse multilingual asr model. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Zhou, T., Wang, S., Liu, J., Yu, Z., & Xiang, Y. (2021). A configurable multilingual model is all you need. *arXiv preprint arXiv:2112.09079*.

Appendices

A English Experiments

Table 6: 10-minute English Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	33.3333	0.482	0.7436	3.4183
200	66.6667	0.2402	0.5833	3.4448
300	100.0	0.0135	0.3881	3.5834
400	133.3333	0.0029	0.3731	3.6324
500	166.6667	0.0019	0.3685	3.6568
600	200.0	0.0014	0.3663	3.6854
700	233.3333	0.0012	0.3649	3.7098
800	266.6667	0.0011	0.3641	3.7241
900	300.0	0.001	0.3637	3.7465
1000	333.3333	0.001	0.3635	3.7424

Table 7: 1-hour English Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	5.1282	0.5827	0.7468	3.4509
200	10.2564	0.3801	0.5781	3.4856
300	15.3846	0.1166	0.2330	3.8872
400	20.5128	0.0469	0.1750	4.1053
500	25.6410	0.0249	0.1637	4.1277
600	30.7692	0.0173	0.1609	4.1297
700	35.8974	0.0119	0.1604	4.1358
800	41.0256	0.0087	0.1607	4.1501
900	46.1538	0.0074	0.1609	4.1460
1000	51.2821	0.0071	0.1610	4.1481

Table 8: 10-hour English Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	0.5556	0.525	0.7431	3.4571
200	1.1111	0.382	0.5645	3.4836
300	1.6667	0.1704	0.2111	4.0237
400	2.2222	0.0953	0.1527	4.1114
500	2.7778	0.0904	0.1404	4.0400
600	3.3333	0.0784	0.1355	4.0482
700	3.8889	0.0793	0.1331	3.9768
800	4.4444	0.0776	0.1318	3.9646
900	5.0	0.0629	0.1310	3.9830
1000	5.5556	0.0746	0.1307	3.9809

Table 9: 10-minute English Training Results for Distil-whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	33.3333	0.5641	0.9641	3.4754
200	66.6667	0.3271	0.7822	3.4652
300	100.0	0.0871	0.5731	3.4530
400	133.3333	0.0149	0.5142	3.4774
500	166.6667	0.0043	0.5051	3.5345
600	200.0	0.0026	0.5030	3.5569
700	233.3333	0.002	0.5020	3.5671
800	266.6667	0.0016	0.5015	3.5773
900	300.0	0.0014	0.5013	3.5936
1000	333.3333	0.0014	0.5012	3.5814

Table 10: 1-hour English Training Results for Distil-whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	5.1282	0.693500	0.974246	3.463177
200	10.2564	0.496900	0.767838	3.475407
300	15.3846	0.284000	0.508848	3.503944
400	20.5128	0.143500	0.353318	4.078762
500	25.6410	0.093200	0.306344	4.323366
600	30.7692	0.071900	0.284102	4.443629
700	35.8974	0.053600	0.271939	4.449744
800	41.0256	0.043400	0.266289	4.464013
900	46.1538	0.038200	0.263742	4.447706
1000	51.2821	0.036900	0.263008	4.466051

Table 11: 10-hour English Training Results for Distil-whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	0.5556	0.651	0.9641	3.4754
200	1.1111	0.5006	0.7651	3.5039
300	1.6667	0.3531	0.5188	3.5121
400	2.2222	0.2176	0.3514	4.0258
500	2.7778	0.1834	0.2878	4.3132
600	3.3333	0.1587	0.2589	4.4049
700	3.8889	0.1553	0.2447	4.5007
800	4.4444	0.1566	0.2370	4.5007
900	5.0	0.1226	0.2332	4.5048
1000	5.5556	0.1533	0.2318	4.4905

B Frisian Experiments

Table 12: 10-minute Frisian Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
50	3.3333	1.6208	1.6201	70.3083
100	6.6667	0.0687	1.4034	58.9699
150	10.0	0.0038	1.3975	56.5104
200	13.3333	0.0019	1.4150	56.2966
250	16.6667	0.0015	1.4198	56.2538

Table 13: 1-hour Frisian Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	1.1236	0.9012	1.0076	50.6327
200	2.2472	0.2217	0.8082	42.3311
300	3.3708	0.0728	0.7689	39.8467
400	4.4944	0.0228	0.7767	38.6526
500	5.6180	0.0099	0.7866	38.7738
600	6.7416	0.0061	0.7909	38.7382

Table 14: 10-hour Frisian Training Results for Whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
100	0.1070	1.0548	1.0200	52.0620
200	0.2139	0.6944	0.7126	39.6222
300	0.3209	0.6024	0.6052	36.0791
400	0.4278	0.4697	0.5303	32.5040
500	0.5348	0.4222	0.4780	30.0766
600	0.6417	0.4075	0.4458	28.4691
700	0.7487	0.3740	0.4151	26.9292
800	0.8556	0.3381	0.3949	25.4678
900	0.9626	0.3235	0.3764	24.8904
1000	1.0695	0.1861	0.3643	23.5716
1100	1.1765	0.1554	0.3608	23.4183
1200	1.2834	0.1639	0.3511	23.0298
1300	1.3904	0.1453	0.3449	22.6591
1400	1.4973	0.1531	0.3419	22.4452
1500	1.6043	0.1299	0.3402	22.4274

Table 15: 10-minute Frisian Training Results for Distil-whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
500	33.3333	0.1154	3.7440	105.5534
1000	66.6667	0.0107	4.0814	88.1911
1500	100.0	0.0	4.1314	90.7432
2000	133.3333	0.0	4.1700	89.8663
2500	166.6667	0.0	4.1878	89.6061
3000	200.0	0.0	4.1946	89.3637

Table 16: 1-hour Frisian Training Results for Distil-whisper-small

Step	Epoch	Training Loss	Validation Loss	WER
500	5.6180	0.2482	1.8089	66.9720
1000	11.2360	0.1076	1.8466	62.2349
1500	16.8539	0.0448	1.9436	59.3548
2000	22.4719	0.0062	1.8986	56.5960
2500	28.0899	0.0016	1.9025	54.4324
3000	33.7079	0.0001	1.9212	54.3005

Table 17: 10-hour Frisian Training Results for Distil-whisper-small

Step	**Epoch**	**Training Loss**	**Validation Loss**	**WER**
500	0.5348	0.9504	1.0939	57.4122
1000	1.0695	0.4656	0.8241	45.7316
1500	1.6043	0.4533	0.7285	41.3474
2000	2.1390	0.1745	0.6875	37.7009
2500	2.6738	0.1701	0.6261	34.7603
3000	3.2086	0.0709	0.6566	33.4415
3500	3.7433	0.0731	0.5880	30.5650
4000	4.2781	0.0234	0.5949	28.8754
4500	4.8128	0.0192	0.5799	27.7063
5000	5.3476	0.0038	0.5755	26.9114