university of groningen

campus fryslân

From Zero-Shot to Fine-Tuned: Linguistic Error Analysis in Frisian ASR with Whisper

Xinchi Li

University of Groningen - Campus Fryslân

From Zero-Shot to Fine-Tuned: Linguistic Error Analysis in Frisian ASR with Whisper

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
Assoc. Prof. Dr. Matt Coler (Voice Technology, University of Groningen) with the second reader being **Assis. Prof. Dr. Joshua Schäuble** (Data Science, University of Groningen)

Xinchi Li (S-5853532)

July 7, 2025

Acknowledgements

There are many people I would like to thank for their support in the completion of this thesis.

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, Matt Coler, for your patient guidance and precise suggestions. From the very beginning, during the proposal stage, you offered valuable feedback with great care. When I encountered difficulties during the experimental phase of my research, you were quick to respond and provided constructive insights that allowed for a more rigorous experimental design. I am also especially grateful for your detailed review of the full manuscript, which greatly contributed to the quality of this thesis.

Secondly, I want to thank Igor Marchenko for your support when I faced technical challenges. Your help made the implementation and running of the experimental models much smoother.

I would also like to thank myself. It was the result of two months of hard work and dedication that brought this thesis to life. Although it may not be the most outstanding thesis, it is a testament to my persistent effort and a representation of what I have learned during my master's studies.

Lastly, I would like to thank my parents. Without your unwavering support, I would never have had the opportunity to study speech technology here. Thank you for giving me the chance to explore more possibilities in my life.

Abstract

Frisian is a low-resource language that shares close linguistic ties with Dutch, German, and English. Automatic Speech Recognition (ASR) projects for Frisian have long faced challenges such as limited availability of speech and transcription data, as well as low model accuracy. This study investigates how to effectively model Frisian using Whisper(small), a multilingual pre-trained model, through cross-lingual transfer learning. This approach leverages Whisper's built-in multilingual tokenizer, eliminating the need for Frisian-specific preprocessing Additionally, we analyze the causes of recognition errors from a linguistic perspective after cross-lingual adaptation.

We selected the Dutch, German, and English configurations of the Whisper model and conducted both zero-shot testing and fine-tuning experiments. The results show that, without fine-tuning, the Word Error Rates (WER) of the models were: Dutch – 90.84%, German – 104.052%, and English – 111.954%. After fine-tuning on Frisian data, the WERs significantly decreased to: Dutch – 5.745%, German – 5.877%, and English – 5.741%. These findings prove the strong potential of cross-lingual transfer learning in Frisian ASR, especially when the source and target languages are closely related and structurally similar. High recognition accuracy was achieved without the need for additional language models or customized tokenizers.

Linguistic analysis of the ASR errors revealed common issues such as language transfer effects, grammatical marker confusion, and phonetic similarity confusions. This study confirms the feasibility and efficiency of using multilingual pre-trained models for transfer learning in low-resource languages and provides insights into error types and future directions for low-resource ASR system development.

keywords:Cross-lingual Transfer Learning,Frisian Speech Recognition,Linguistic Error Analysis

Contents

1 Introduction

With the significant advancements in Automatic Speech Recognition (ASR) for high-resource languages(Graves, Mohamed, & Hinton, 2013), extending these developments to low-resource languages has become a key research direction in the field of speech technology (Besacier, Barnard, Karpov, & Schultz, 2014). Low-resource languages often face challenges such as limited available speech data, lack of dedicated tokenizers, and low-accuracy language models(Babu et al., 2021). Frisian is a typical example of such a low-resource language, as it has a relatively small number of speakers, limited data resources, and insufficient support from speech recognition tools. It is primarily spoken in the province of Friesland in the northern part of the Netherlands. Enhancing ASR technology for low-resource languages like Frisian can help preserve these valuable linguistic resources by improving their accessibility and increasing their visibility, thereby contributing to their protection and continued use.

The development of ASR systems for low-resource languages faces numerous obstacles, such as data scarcity, lack of tool support, and high development costs. For languages like Frisian, which suffer from insufficient linguistic resources and limited usage scenarios—leading to low model return on investment—it is often unrealistic to build ASR systems from scratch. In recent years, multilingual pre-trained models have opened new possibilities for modeling low-resource languages(Babu et al., 2021). However, existing studies have rarely examined the effectiveness of transfer learning between linguistically related languages. Prior research has predominantly examined transfer learning between either typologically distant languages or resource-rich pairs, neglecting the role of linguistic relatedness in transfer effectiveness(Bansal, Kamper, Livescu, Lopez, & Goldwater, 2018; Wang, Pino, & Gu, 2020; Yadav & Sitaram, 2022) . Consequently, empirical studies on transfer learning between structurally similar yet resource-imbalanced languages remain scarce and underexplored in the academic community.

From a genealogical perspective, Frisian belongs to the West Germanic branch of the Indo-European language family and shares close linguistic ties with Dutch, German, and especially Old English. These connections are evident in phonology, vocabulary, and other linguistic features(Ringe & Taylor, 2014). Such structural proximity offers promising potential for transfer learning: conducting transfer learning from three linguistically related high-resource languages provides a more targeted and efficient approach compared to transferring between typologically distant languages. This can significantly reduce the time and material costs of developing ASR systems for Frisian, thereby supporting the preservation and revitalization of this endangered language. While English shares the closest historical genealogical relationship with Frisian, modern Dutch and Frisian exhibit greater structural similarity due to centuries of contact and convergence(Jong, 2015). This creates competing predictions: historical relatedness favors English, while synchronic structural similarity favors Dutch for transfer learning effectiveness.

In recent years, multilingual pre-trained models have introduced new possibilities for modeling low-resource languages(Babu et al., 2021). The Whisper model, trained and fine-tuned on 680,000 hours of labeled data across more than 95 languages, has demonstrated strong cross-lingual generalization capabilities. Its technical approach and detailed architecture will be introduced in Section 3.1 of this thesis. However, due to its multilingual design and robust generalization ability, this study does not require any specialized tokenizers or complex preprocessing procedures. Whisper's multilingual pretraining and minimal preprocessing requirements make it suitable for low-resource language applications.

Based on the linguistic relationship between Frisian and other languages, as well as the current state of research in the field for low-resource languages(Babu et al., 2021; Choe et al., 2022; Ringe & Taylor, 2014; Żelasko, Moro-Velázquez, Hasegawa-Johnson, Scharenborg, & Dehak, 2020) , this study postulates the following research questions:

1.Without using a dedicated tokenizer or language-specific preprocessing, Will the performance of Frisian automatic speech recognition be significantly improved by applying transfer learning and fine-tuning using zero-shot Whisper-small models pre-trained on Dutch, German, and English?And which language will perform best?

2.Can we explain the prediction errors generated by these models after fine-tuning from a linguistic perspective reasonably?

According to the multilingual generalization capability of the Whisper model and relevant studies in the field(Bansal et al., 2018; Choe et al., 2022; Ringe & Taylor, 2014; Żelasko et al., 2020) , this study proposes the following hypotheses:

Hypothesis 1: Fine-tuning the Dutch, German, and English-based zero-shot Whisper-small models on Frisian data will lead to a significant improvement in ASR performance. Based on the structural similarity between Dutch and Frisian (Jong, 2015), the Dutch-based model is expected to achieve the best performance.

Hypothesis 2: The differences in model performance across source languages will be influenced by linguistic factors, such as the lexical and syntactic differences between the source language and Frisian.Therefore, the richness and differences in language structure may lead to Compound Word Errors, where the model struggles with the correct segmentation of complex words.

Against the backdrop of limited research on transfer learning between structurally similar but resource-imbalanced languages, this study fills an important gap through empirical investigation. By using Whisper-small models configured for three languages closely related to Frisian—Dutch, German, and English—this research systematically evaluates the feasibility of transfer learning for Frisian ASR without relying on customized tokenizers or special preprocessing. Furthermore, by analyzing prediction errors from a linguistic perspective, the study reveals how linguistic differences between source and target languages can impact transfer learning outcomes. These insights contribute to the refinement of transfer strategies in multilingual ASR models and provide a theoretical basis for language selection in low-resource settings. In addition, this work offers a replicable framework for studying other low-resource languages with similar linguistic profiles.

1.1 Thesis Outline

The structure of this thesis is organized as follows. Having presented the research background and motivation, stated the research questions and hypotheses, and given a brief overview of the thesis structure, the next section is dedicated to the Literature Review. This section includes reviews related work on ASR for low-resource languages and cross-lingual transfer learning, including characteristics of the Frisian language and common types of ASR errors. Section 3, Methodology, describes the model selection, dataset construction, experimental procedure, evaluation metrics, and ethical considerations, forming a complete experimental design framework. The following section, 4 Results, shows the transfer learning results of the three language models on Frisian speech recognition, with a detailed comparison and analysis of prediction errors. 5 Discussion interprets the experimental findings, evaluates the research hypotheses, summarizes the linguistic reasons behind model performance differences, and discusses the study's limitations and future directions. Finally, 6 Conclusion summarizes the research goals and findings, highlights the study's contribution to low-resource ASR and cross-lingual transfer learning, and looks ahead to potential real-world applications.

2 Literature Review

This section provides an overall review of research on automatic speech recognition (ASR) for low-resource languages and cross-lingual transfer learning. Through a deep and critical analysis of the literature in these areas, we can not only have a better understanding of speech recognition technologies for low-resource languages using transfer learning, but also find out other potential directions for future research.

The structure of this chapter is as follows: First, a review of existing ASR studies on low-resource languages; Second, an introduction to cross-lingual transfer learning; Third, a comparison of cross-lingual transfer methods, and make explicit the gap analysis connecting to the RQ; Fourth, a description of previous ASR approaches for Frisian; Fifth, a description of Frisian and its related languages; And finally, a discussion of common ASR errors from a linguistic perspective.

2.1 Low-Resource Languages ASR

ASR for low-resource languages has gone through a long period of research and development. This study employed a systematic literature review methodology. Using Google Scholar with the search terms “low-resource language” AND “survey” AND “automatic speech recognition.” No restrictions were placed on publication date or language in order to ensure comprehensive coverage of relevant research. Inclusion criteria required that studies focus on ASR for low-resource languages, include either a review or empirical analysis, and provide access to the full text. Exclusion criteria eliminated works that were not directly related to ASR, offered only superficial discussions of the topic, or lacked practical case studies. Low-resource ASR development has evolved through four distinct stages: HMM-GMM systems, deep learning approaches, end-to-end methods, and self-supervised learning(Yadav & Sitaram, 2022).

The first stage involves the traditional HMM-GMM model. The Hidden Markov Model – Gaussian Mixture Model (HMM-GMM) system was a classic approach in early ASR research. In this model, speech signals such as Mel-Frequency Cepstral Coefficients (MFCCs) are first extracted as acoustic feature sequences. These features are then input into the HMM-GMM system for modeling and recognition. Each basic speech unit is typically represented by a five-state HMM, with the first and last states being non-emitting. The observation probabilities of each state are modeled using GMMs, which represent the distribution of acoustic features as a weighted sum of several Gaussian distributions, allowing the system to predict feature patterns (Pujol, Pol, Nadeu, Hagen, & Bourlard, 2005).

The second stage is characterized by the use of deep learning methods such as LSTM. LSTM stands for Long Short-Term Memory, a method that introduces memory cells and gating mechanisms to effectively learn long-term dependencies. This significantly improves performance in both speech enhancement and ASR. However, it also requires large amounts of accurately labeled data(Weninger et al., 2015).

The third stage is the development of end-to-end methods. These approaches combine the acoustic, pronunciation, and language models of traditional ASR systems into a single neural network.

This greatly reduces the number of parameters and simplifies the whole process, lowering the cost of data annotation(Li et al., 2020). One example is Connectionist Temporal Classification (CTC), which does not require the output sequence to align exactly with the input frames. CTC introduces repeated labels and blank symbols to enable training without alignment. During decoding, a greedy decoding strategy is used to select high-confidence labels that are not blanks, further simplifying the process(Li et al., 2020).

The fourth and most recent stage is self-supervised learning (SSL). SSL makes use of large amounts of unlabeled data for pretraining and has shown excellent performance in tasks such as speech recognition and speaker verification(Liu et al., 2023). For example, wav2vec 2.0 uses a convolutional neural network (CNN) to extract speech features, which are then modeled using a Transformer to capture contextual information. Through contrastive prediction tasks and other self-supervised objectives, the model learns useful speech representations without the need for manual transcription(Kozhirbayev, 2023) .

Each developmental stage has progressively addressed the challenges of data scarcity and acoustic modeling difficulties in low-resource language ASR. In particular, the most recent stage—self-supervised learning—has greatly reduced the reliance on labeled data by enabling pretraining on large number of unlabeled speech. This has significantly improved the modeling capacity for low-resource languages and broadened the path for the application of ASR in multilingual situations.

2.2 Cross-lingual transfer learning

Cross-lingual transfer learning is an effective approach to improve the performance of automatic speech recognition (ASR) systems for low-resource languages. The main idea of this method is to develop the modeling ability of ASR systems in low-resource target languages by transferring linguistic knowledge learned from high-resource source languages. This strategy is especially beneficial when labeled data of the target language is limited.

This study adopted a systematic literature search methodology, using Google Scholar with the keywords “cross-lingual transfer” AND “automatic speech recognition.” No restrictions were placed on publication date or language to ensure a broad scope. The inclusion criteria required that studies focus on cross-lingual transfer methods in ASR, contain either a review component or empirical analysis, and be accessible in full text. Studies were excluded if they did not involve actual cross-lingual transfer applications, were not directly related to ASR, or lacked transparency in their research methodology. We can find various studies on this topic. Some research suggests that cross-lingual transfer learning involves transferring the modeling abilities of high-resource language systems to those for low-resource languages, and that speech translation can play a supportive role in this transfer process(Wang et al., 2020). Other studies point out that transfer learning can be used not only through a ”pretraining and fine-tuning” approach, but also during multilingual training. This allows low-resource languages to be benefited from high-resource languages that are linguistically similar(Yadav & Sitaram, 2022).

In summary, cross-lingual transfer learning can be seen as a method of transferring mature modeling knowledge from high-resource languages to low-resource ones. It is currently one of the most effective ways for improving ASR performance in low-resource language settings.

2.3 Comparison of cross-lingual transfer methods

Cross-lingual transfer learning can be applied in several ways.The most common approach is the ”pretraining and fine-tuning” strategy. For example, a Tacotron 2 model can be pretrained using a large English dataset, which is a high-resource language. By applying a Phoneme Transformation Network to share phonemes across languages, the pretrained model can then be fine-tuned using only 15 to 30 minutes of data from low-resource languages such as Chinese, German, and French. Even with such limited labeled data, the model is still able to generate fluent speech, showing the effectiveness of this transfer learning method(Tu, Chen, Yeh, & Lee, 2019). However, this approach has not systematically explored the effectiveness of transfer learning between linguistically related languages. For instance, Dutch, German, English, and Frisian all belong to the Germanic language family and theoretically share strong transfer potential. Yet, there is currently a lack of systematic empirical research focusing on such closely related language pairs.

Another approach is “multilingual joint training”. This method focuses on training a single model on multiple languages at the same time, using the data advantage of high-resource languages to improve the performance for low-resource ones. In practice, part of the model—especially the encoder—is shared with all languages, allowing it to learn common acoustic features. At the same time, each language remains its own output layer to handle specific pronunciation differences(Yadav & Sitaram, 2022). This method is especially useful for languages with similar linguistic structures and showing great potential in improving ASR models for low-resource languages.

Finally, some studies apply auxiliary tasks for transfer learning, such as speech translation. As we mentioned before(Wang et al., 2020), they trained ASR and speech translation (ST) tasks together in a multitask learning setup. This allows the model to learn from the translation task, even when the target language has very limited data. However, such approaches rely on the availability of bilingual or multilingual datasets with paired speech and translation texts, which is nearly unfeasible for low-resource languages like Frisian.

In conclusion, cross-lingual transfer learning has become an important way for improving ASR performance in low-resource languages. Approaches such as pretraining and fine-tuning, multilingual joint training, and the use of auxiliary tasks all overcome the challenge of data scarcity from different angles. However, existing studies rarely focus on transfer learning between linguistically similar languages that different in resource availability. Therefore, this study proposes to explore cross-lingual transfer learning using high-resource languages that are closely related to Frisian—Dutch, German, and English—in order to fill this gap in the current research.

2.4 Previous ASR approaches for Frisian

To better understand the differences and innovations of the methods used in this study compared to previous research in related fields, the following table presents a comparison between earlier studies and the current research.

Study	Languages	Method	Error analysis	Dataset	Best WER
(Khurana, Laurent, & Glass, 2022)	Monolingual English wav2vec-2.0 → Frisian	DUST self-training	none	Multilingual corpora incl. Frisian	comparable to XLS-R (exact WER not reported)
(Bălan, 2023)	Multilingual ↦ Frisian	XLS-R fine-tuning	none	Common Voice 12.0	15.99%
(Amooie et al., 2025)	Dutch + German + English → Frisian	XLS-R multilingual fine-tuning + LID	yes (standard vs dialectal WER)	Common Voice 17.0	13.1%

Table 1: Comparison of prior Frisian ASR studies

In summary, previous research has primarily focused on fine-tuning large-scale pre-trained models on Frisian, but often based on single-source or linguistically unrelated language choices, with a general lack of in-depth analysis of interlinguistic relationships. For instance, the first study adapted Frisian using a monolingual English model via self-training but did not explicitly report WER results(Khurana et al., 2022) ; the second one employed multilingual fine-tuning but did not provide a detailed analysis of recognition errors(Bălan, 2023); the third one remains the only study to differentiate to some extent the effects of transfer from multiple related languages, yet their analysis focused primarily on model performance evaluation rather than systematic linguistic error analysis(Amooie et al., 2025).

In contrast, this study is the first to systematically compare ASR transfer learning outcomes from three Germanic source languages—Dutch, German, and English—into Frisian, incorporating in-depth linguistic error analysis. This approach not only helps identify how the source language impacts ASR performance for the low-resource target language, but also offers linguistic motivation and empirical evidence for future cross-lingual ASR system design.

2.5 Frisian and Its Related Languages

This study employed a systematic literature review methodology by searching Google Scholar with the keywords "Frisian and its related languages" AND "Frisian languages."And also used the Taalportaal website to search for knowledge about frisian. No restrictions were applied regarding publication date or language. The inclusion criteria required that studies focus on Frisian or its dialects/related languages, particularly in the domains of linguistic descriptions, historical development, language resources, or speech recognition, and that full texts be accessible. Studies were excluded if they only mentioned Frisian indirectly or were unrelated to language technology or historical development. Ultimately, several highly relevant studies were selected to serve as the theoretical foundation for this research.

Frisian is a branch of the West Germanic languages, specifically classified under the Coastal West Germanic subgroup. It is closely related to English, while Dutch and German belong to the Continental West Germanic subgroup. Originally, Frisian was spoken in the northern coastal regions along the North Sea in what is now the Netherlands and Germany. Traditionally, the language is divided into three main dialects: West Frisian, East Frisian (Saterland dialect), and North Frisian. Although these dialects share certain features—such as the presence of two classes of weak verbs—they have developed along different paths due to significant geographical and linguistic differences in their environments. As a result, mutual intelligibility among the three has largely disappeared(de Graaf, 2016).

This study focuses on West Frisian, which holds a recognized status in Dutch society and has been officially designated as the second official language of the Netherlands(Winter, 2022). West Frisian consists of three main dialects: Clay Frisian (Klaaifrysk), Forest Frisian (Wâldfrysk), and Southwestern Frisian (Súdwesthoeksk). Speakers of Clay Frisian tend to speak at a slower pace and often produce longer vowels, which frequently result in diphthongs. In contrast, Forest Frisian speakers generally speak more quickly and exhibit a distinct phonological process known as ”breaking”—This involves the alternation of centring diphthongs—such as/iə/ and /jə/---into glide plus vowel sequences like [jɪ] or [jɛ] in complex forms such as plurals, diminutives, and compounds. This phenomenon is referred to as Modern Frisian Breaking(Visser, 2015). Additionally, Forest Frisian speakers are known for their short and clear pronunciation of personal pronouns—free morphemes used to refer to people, animals, objects, substances, or abstract concepts(Dyk, undefined). The third dialect, Southwestern Frisian, is spoken in the southwestern corner of Friesland. It differs phonologically from the other two varieties. While it does not exhibit the breaking feature, it achieves a similar morphological function by introducing an extra phoneme that is absent in the other dialects. The standard variety of West Frisian is primarily based on Clay Frisian, though it omits the drawn-out vowel pronunciation characteristic of that dialect(Jong, 2015).

Modern Frisian still retains several linguistic features similar to English, which is its closest relative within the Germanic family(Van Heuven & Kirsner, 2004). However, over time, Frisian has been heavily influenced by Dutch, leading to increasing similarities between the two languages. In Germany, Low German has largely replaced Frisian in the regions where East and North Frisian were historically spoken, though Frisian substratum influences remain in the local dialects(de Graaf, 2016).

In the field of Frisian language research, the Fryske Akademy has played an important role. Since its appearance, the institute has published over one thousand books in various languages, about one-third of which are in Frisian. The rest are mainly in Dutch, English, and German(Jong, 2015), highlighting the strong connections between Frisian and these three languages. Based on this linguistic and historical relationship, this study selects zero-shot models trained on Dutch, German, and English as the foundation for fine-tuning the Frisian ASR model.

2.6 Linguistic Analysis of ASR Errors

This study employed a systematic literature review methodology by searching Google Scholar using the keywords “linguistic analysis of ASR errors” AND “reason of ASR errors.” No restrictions were applied regarding publication date or language. The inclusion criteria required that the studies focus on linguistic explanations of ASR errors and provide clear classifications of error types, with full texts available. Studies were excluded if they did not explicitly analyze the causes of recognition errors or only discussed technical issues. Ultimately, several representative papers were selected to support the analysis and discussion of ASR error patterns in this research.

Sometimes, ASR models generate prediction errors. These errors can be understood as resulting from ambiguous speech regions, where the audio input and/or its context create confusion, ultimately leading to discrepancies between the predicted and reference transcriptions. Such discrepancies may arise from two main sources: the first is model bias, caused by simplified or imperfect ASR architectures; the second is linguistic bias, stemming from the inherent ambiguity of natural language(Adda-Decker, Vasilescu, Snoeren, Yahia, & Lamel, 2011).

Linguistic research on ASR errors has identified several linguistic factors that frequently contribute to recognition failures. These factors highlight the complex interaction between speech signal processing and the structural characteristics of natural language.

1. Compound Word Errors

Compound word errors appear when ASR systems misinterpret or incorrectly segment compound structures, which are particularly prevalent in morphologically rich languages such as Sanskrit, German, and Dutch. These languages often contain long compound words with multiple morphemes, and incorrect boundary detection can result in recognition errors or unintended word formations(Kumar et al., 2022).

2. Syncretism

Syncretism refers to the phenomenon where a single word form is used for multiple grammatical functions. This ambiguity can hinder the ASR system's ability to assign the correct syntactic or morphological role to the recognized word, especially in languages with high levels of morphological inflection(Kumar et al., 2022).

3. Homophony

Homophony, the existence of different words sharing identical pronunciations, also contributes to recognition errors. When contextual information is insufficient, ASR models may fail to disambiguate homophones correctly, resulting in semantically or syntactically inappropriate outputs (Kumar et al., 2022).

4. Collocational variation

Collocational variation occurs when phonetically similar expressions differ in terms of their typical usage patterns or stylistic registers. Such variation can introduce pragmatic ambiguity, which challenges the model's contextual inference capabilities. Misinterpretation of collocational patterns may lead to errors, especially in informal or conversational speech.

It is important to note that syncretism, homophony, and collocational variation often cause both human listeners and ASR systems to understand unclear or degraded speech in different ways, which increases the chance of recognition errors(Adda-Decker et al., 2011).

5. Liaison and Phonetic Ambiguity
Spoken language frequently exhibits coarticulation, liaison, segmental reduction, or sound deletion. These phonetic phenomena tend to obscure word and morpheme boundaries, leading to blurred or collapsed sound segments. ASR systems may misdecode such regions due to the lack of clear acoustic boundaries, particularly in spontaneous or fluent speech(Adda-Decker et al., 2011).

6. Ambiguous speech regions
Ambiguous speech regions refer to portions of audio where either the acoustic signal or the linguistic context is unclear. These regions are difficult to transcribe accurately even for human annotators, and they often result in inconsistent or erroneous ASR output(Adda-Decker et al., 2011).

Based on these identified sources of error, this study uses a linguistic approach to analyze the ASR outputs of models that were first trained on Dutch, German, and English, and then fine-tuned on Frisian data. By organizing and explaining the recognition errors using linguistic theory, this research aims to give a clear understanding of error patterns and offer helpful ideas for improving cross-lingual ASR systems in the future.

3 Methodology

This chapter provides a detailed overview of the methodological framework adopted in this study to address the research questions and test the proposed hypotheses. The structure is designed to ensure that the methods used are systematic, repeatable, and scientifically sound. It covers all important parts, including model selection, dataset construction, experimental design, evaluation metrics, as well as ethical considerations and data privacy.

Specifically, the chapter is organized into the following parts: first, it introduces the selected models and the reasons for their choice; second, it presents the datasets used in the study; third, it describes the experimental setup and parameter configurations; fourth, it explains the evaluation metrics for measuring model performance; and finally, it discusses ethical compliance.

3.1 Model

Model selection directly impacts the validity and reproducibility of experimental results in cross-lingual transfer learning studies. The right model allows for the effective use of available resources and data, and helps produce scientifically reliable results.The Whisper model, developed by OpenAI, is a pretrained model designed for ASR and speech translation. More specifically, it is a Transformer-based encoder-decoder model, also known as a sequence-to-sequence model. Whisper was trained on 680,000 hours of labeled speech data, with large-scale weak supervision used for annotation. The Whisper models were trained either on English-only data or on multilingual data. The English models are optimized for ASR tasks only, while the multilingual models are trained for both speech recognition and speech translation. In ASR tasks, the model predicts transcriptions in the same language as the input audio. In speech translation tasks, it predicts translated text in a different language than the audio input(Radford et al., 2023).

Because it was trained on 680,000 hours of labeled data, the Whisper model has strong generalization capabilities and can perform well across a wide range of datasets and applications without the need for fine-tuning(Radford et al., 2023). This makes it highly valuable for cross-lingual transfer learning. Therefore, to investigate the performance of Frisian data in transfer learning using Dutch, German, and English based zero-shot models, Whisper is a very suitable choice for this study.

The Whisper model is available in five different configurations of increasing size: tiny, base, small, medium, and large. The first to fourth versions—tiny to medium—are available in both English-only and multilingual training variants, while the largest model is available only in a multilingual version(Radford et al., 2023). In this study, I selected the whisper-small model. The reasons are as follows: First, the small version has fewer parameters, making it a more practical choice given the limited computing resources and time available for experimentation. Second, despite its smaller size, whisper-small remains the strong generalization ability of the Whisper architecture. The reduction in size does not significantly effect performance, especially in low-resource settings. Finally, for a language like Frisian, where annotated data is limited, using a smaller model helps reduce the risk of overfitting during fine-tuning. As noted in related literature, it is recommended to limit fine-tuning steps to around 5,000 for Whisper-small, so this study also capped the maximum number of steps at 5,000(de Zuazo, Navas, Saratxaga, & Rioja, 2025).

To further explain how Whisper works, the next section will introduce its main architecture and training tasks.

3.1.1 Model architecture

Because the focus is on exploring the power of large-scale supervised pretraining in speech recognition, Whisper uses a standard and well-tested model architecture—the encoder-decoder Transformer, which is known to scale well and perform reliably(Radford et al., 2023).The core architecture of Whisper consists of four main parts: input processing, encoder, decoder, and multitask output.

In the preprocessing stage, all audio data is resampled to 16,000 Hz, and 80-channel log-Mel spectrograms are extracted using a 25 ms window and a 10 ms hop size. During feature normalization, the model scales the input spectrograms across the entire training set so that the values are between [-1, 1], with a mean close to 0.

In the encoder part, the spectrograms are first processed by two 1D convolutional layers. These layers use a kernel size of 3 and the GELU activation function. The second convolutional layer reduces the time resolution, making the sequence shorter. Then, sinusoidal positional encodings are added to the output and sent into multiple Transformer encoder layers. These layers use pre-activation residual connections, and layer normalization is applied at the end.

In the decoder stage, the model uses learnable positional encodings and shares the input and output embedding weights. In every model version, the encoder and decoder have a symmetrical structure in terms of depth and width. The full structure is shown in Figure 1.

In the Multitask Output stage, Whisper encodes the task type and context as a sequence of special instruction tokens, which are given to the decoder as input. The process works as follows: Each audio segment starts with a `<|startoftranscript|>` token, which marks the beginning of transcription. Then, the model detects the spoken language and uses a language token to indicate it (supporting up to 99 languages). After that, the model receives a task token, such as `<|transcribe|>` for transcription or `<|translate|>` for translation. Tokens like `<|notimestamps|>` or `<|timestamp|>` can also be added to control whether timestamps are included in the output. With these tokens, the model understands what task to perform and what format the output should follow, and then starts generating the corresponding text output (Radford et al., 2023).

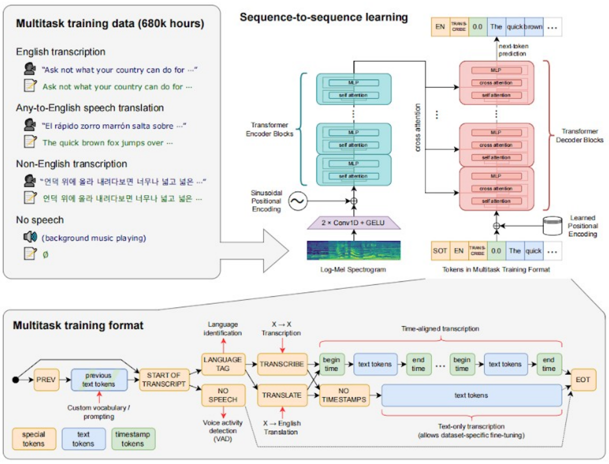

Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
...
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
cross attention
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Learned Positional Encoding
Log-Mel Spectrogram
SOT
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
NO TIMESTAMPS
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 1: The overview of the whisper model architecture. Reprinted from(Radford et al., 2023)

3.1.2 Training Objectives

Whisper jointly learns multiple tasks (Radford et al., 2023): ASR, speech translation, language identification, and voice activity detection (VAD).

Each input example consists of a 30-second audio segment, which is first converted into an 80-dimensional log-Mel spectrogram, and then a sequence of special tokens is added to specify the task type, target language, and timestamp format. These tokens act as context prompts for the decoder. During training, the model uses an autoregressive method. The decoder generates the target text sequence step by step, based on the given input. The training objective is to minimize the cross-entropy loss between the predicted token sequence and the target token sequence. The overall loss function L can be written as:

$$L = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, \mathbf{X})$$

Here, y_t represents the target token at time step t, $y_{<t}$ refers to the sequence of previously generated tokens, and $\mathbf{X}$ is the feature vector output from the audio encoder.

To support multitask learning, the model is trained on a mix of task data, including examples for transcription, translation, and voice activity detection. Each training example is marked with task control tokens (such as `<|transcribe|>`, `<|translate|>`,`<|notimestamps|>`) and language tokens (such as `<|en|>`) to indicate the current task and language setting. In this way, Whisper is able to handle all tasks without using multi-head or branched architectures.

Unlike self-supervised models that often rely on complex objectives like contrastive learning or masked prediction, Whisper uses a standard cross-entropy loss function based on BPE tokens, allowing for a simple and efficient supervised training process.

3.2 Dataset

The Common Voice corpus is a large-scale multilingual speech dataset with transcriptions, designed to support research and development in speech technologies. While it is primarily aimed at ASR tasks, it is also applicable to other areas such as language identification. In the public domain, Common Voice is currently one of the largest and most diverse resources in terms of the number of languages and total recording time(Ardila et al., 2019).

This study uses the Frisian subset of the Common Voice Corpus 21.0. This version contains 4.28 GB of data, totaling 70 hours of speech. The dataset used in this research consists of the training sets, test sets and `other.tsv` file.The `other.tsv` file contains some audio recordings that have not yet been fully validated. However, since Frisian is a low-resource language with limited available data, and the content in `other.tsv` has not been entirely disqualified, it was included in the training process to enhance model performance and increase the amount of training data.

3.3 Experimental Setup

3.3.1 Data Preprocessing

For the Frisian subset of the Common Voice dataset, this study first conducted a careful quality check of the recordings before applying any preprocessing steps. This was done to ensure that the audio clips were free from background noise or other disturbances. In addition, the spoken content was compared with the transcription to eliminate any mismatches or alignment errors. After this review, the study removed all punctuations from the transcriptions and converted all letters to lowercase. This step was taken to avoid any wrong in the experimental results that might be caused by differences in case sensitivity or punctuation marks.

3.3.2 Experiment Design

The experimental design of this study takes the Dutch zero-shot model as an example. First, the Whisper model is set to Dutch as the target language. Then, the model is used to directly predict on the Frisian validation set without any fine-tuning, in order to obtain the initial WER. The predicted output and the reference transcription are printed for each sample to help identify and understand the sources of error. Next, transfer learning is applied by fine-tuning the model on the Frisian training set along with the other subset. After fine-tuning, the model is evaluated again on the validation set, and the updated WER is measured. As before, the predictions and reference transcriptions are printed side by side to support subsequent linguistic analysis of the ASR errors.

Since the English model yielded the best performance among the three zero-shot models, three different random seeds were applied to the English model in the fine-tuning phase to enhance reproducibility and assess result stability. The average WER and standard deviation across these seeds were calculated to provide a more reliable estimate of model performance.

3.3.3 Hyperparameters Setting

After multiple adjustments, this study adopted the most effective set of parameters:

Learning rate: 1e-5

Max steps: 5000

Evaluation and save steps: 250

Per-device training batch size: 16

Random seeds:42,96,132

These settings were selected to balance training efficiency and model performance during fine-tuning.

3.3.4 Evaluation

The evaluation metric used in this study is WER (Word Error Rate). WER is one of the most commonly used performance metrics in speech recognition. It measures the difference between the model's output and the reference text. A lower WER means the recognition is more accurate. The formula for WER is as follows:

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

Here, S stands for substitutions (the number of incorrectly replaced words),D stands for deletions (the number of missing words),I stands for insertions (the number of extra words), and N is the total number of words in the reference text (the ground truth).

Using this evaluation metric, it is possible to clearly compare the accuracy before and after fine-tuning, and to show the actual difference between the predicted text and the reference transcription in numerical form.

3.4 Ethical Considerations

For this experimental study, it is important not only to examine the effectiveness of transfer learning and analyze the experimental results, but also to consider the ethical issues, potential risks and reflection on ethical implications closely related to this research.

3.4.1 Data

The data used in this study comes from the Mozilla Common Voice project. These datasets are open and continuously updated, and are released under the CC0 1.0 Public Domain Dedication, which means they can be freely used, modified, and distributed without payment or the need for additional permission. All speech data was voluntarily recorded and uploaded by participants through the Common Voice platform. The participants explicitly agreed to make their speech data publicly available. The accompanying metadata (such as age and gender) is also provided in an anonymous form.

Therefore, the use of the Mozilla Common Voice dataset in this research is ethically compliant.

3.4.2 Evaluation Metrics

This study uses WER as the evaluation metric, which is a standard and objective measure in the field of speech recognition. No evaluation methods involving human subjectivity were used. The analysis of error causes is also fully based on the defined categories of errors and their actual behavior in the outputs. No subjective judgment is involved in the evaluation process, which helps to eliminate potential ethical concerns.

3.4.3 Transparency and Replicability

All the models used in this study are publicly available on Hugging Face, with detailed `README.md` files provided to ensure reproducibility of the experiments. Although there may be slight variations in results due to hardware differences or randomness, the overall outcomes remain consistent with the described methodology and the main experimental findings.

In summary, this chapter has provided a detailed description of the model selection, dataset usage, experimental setup, evaluation metrics, and ethical considerations. The next chapter will present the results of the experiments, showing the effectiveness of transfer learning across different language models and analyzing the causes of recognition errors.

3.4.4 Reflection on broader ethical implications

In this study, Frisian, as a low-resource language, involves clear ethical dimensions in its language modeling. Therefore, the research was conducted with careful attention to avoiding experimental setups that could compromise the accuracy of results. The study also assumes the responsibility of language preservation by ensuring that inaccurate modeling outcomes are not used in ways that could negatively impact the development of the language. Additionally, the selected Common Voice corpus includes a broad demographic range, covering both genders and speakers aged 20 to 69, thus minimizing the risk of speaker variability affecting error types and helping to mitigate potential ethical risks in the experiment.

3.5 Demonstrator

The interactive demonstrator shows the fine-tuned models performing low-resource ASR using a Whisper-small model. The models have been trained on small datasets and evaluated WER. Audio inputs can be downsampled to 16kHz, and predictions can be compared with reference text, which can show the model performance.

Features:

1.Shows transcription outputs from the fine-tuned Whisper model.
2.Calculates WER for performance evaluation.
3.Allows switching between different language settings.

Technical implementation:

1.Whisper models were loaded and run using the Hugging Face Transformers library.
2.PyTorch and NumPy were used for model processing and data manipulation.
3.Audio was downsampled to 16kHz using scipy.signal.resample.

4.The evaluate library was employed to compute Word Error Rate (WER).

5.Three random seeds were set to ensure reproducibility of the results. The demonstrator code is avaialbe at Xinchi0824/thesis · Hugging Face

4 Results

This chapter presents the experimental results of the models designed to investigate the transfer learning performance after fine-tuning Dutch, English, and German zero-shot models with Frisian data. The main focus of this chapter is to compare the performance of the three language models after transfer learning, using WER as the primary evaluation metric.

Additionally, the predicted texts after fine-tuning are compared with the reference transcriptions for each model. Based on these comparisons, the errors are classified and analyzed one by one from a linguistic perspective, in order to understand the underlying causes of recognition mistakes.

4.1 Transfer learning

The transfer learning experiments aim to evaluate the performance of the three language models after fine-tuning. The model results are summarized in Table 2.

Model	WER (before fine-tuning)	WER (after fine-tuning)
Dutch	90.842%	5.745%
German	104.052%	5.877%
English	111.954%	5.741%
English-42	111.954%	6.123%
English-96	111.954%	6.019%
English-132	111.954%	6.494%

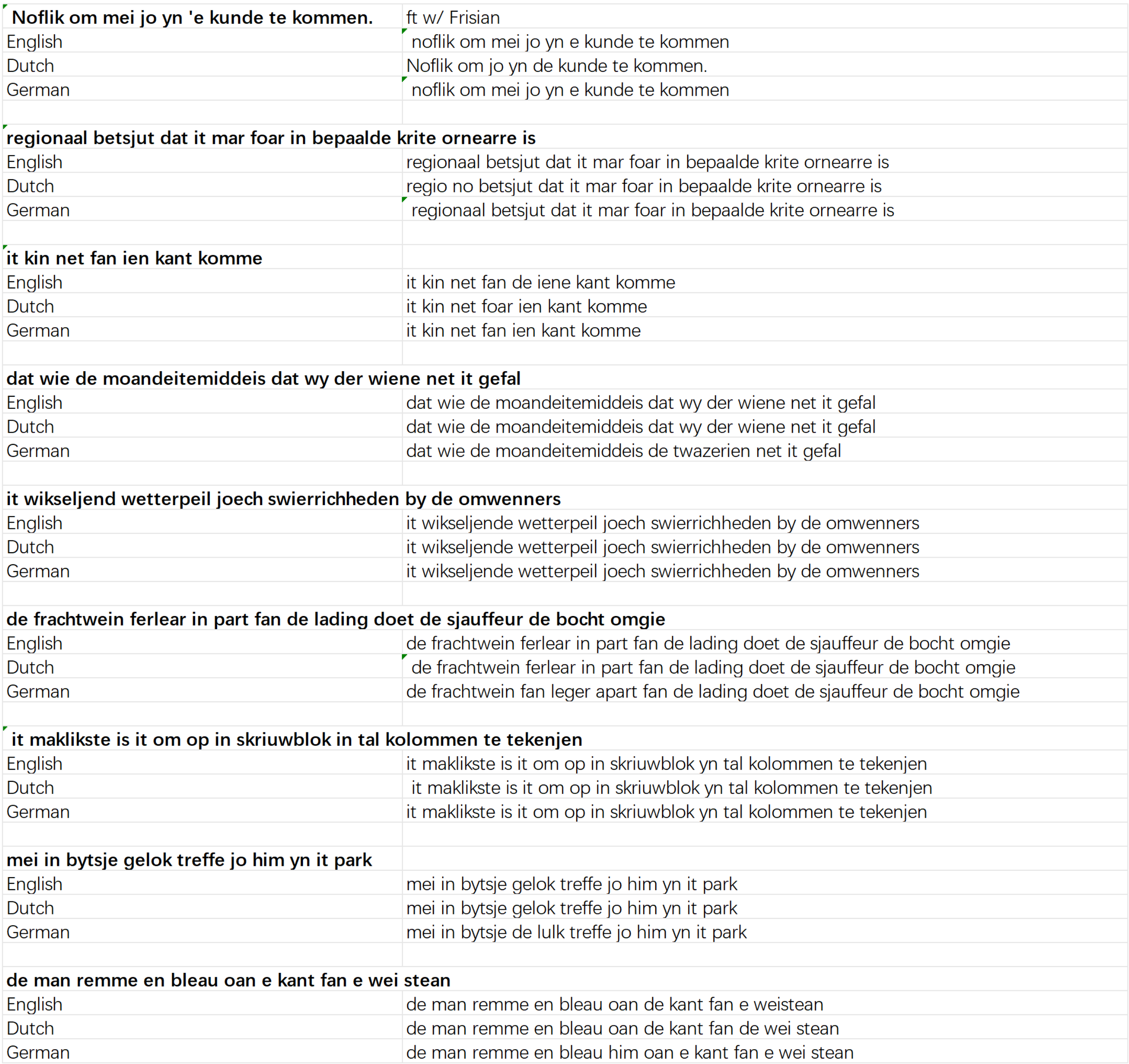
Table 2

100% WER occurs when insertions+substitutions+deletions exceed reference length, common in severe language mismatch. For example: The zero-shot results reveal severe language mismatch, with German and English models producing WER values exceeding 100%. This occurs when the combined errors (insertions, substitutions, and deletions) surpass the total words in the reference transcription—a common phenomenon when models trained on one language attempt to decode acoustically similar but linguistically distinct speech. For instance, the English model likely attempted to force Frisian phonemes into English word patterns, creating numerous spurious insertions. The Dutch model's lower initial WER reflects its closer linguistic relationship to Frisian, though still indicates fundamental recognition failure.

The experimental results show that transfer learning from related languages can indeed improve the performance of ASR models for low-resource languages, demonstrating substantial performance improvements for Frisian ASR. The English-based model achieved the lowest WER (5.741%), outperforming Dutch (5.745%) and German (5.877%) models. To increase the scientific rigor and reproducibility of the experiment, and considering computational limitations, this study conducted three experiments using different random seeds only for the best-performing English model. The results, as shown in the figure above, show that the average WER across the three runs is 6.212%, which is higher than the single-run other languages models without random seed setting, but still below 10%, indicating consistently good performance. The standard deviation is 0.2456%, which is relatively small, suggesting that the English model is not sensitive to random seed changes. This confirms that the training process is stable and reproducible. The reasons why the English model performed best among the three language models without random seed settings will be further discussed in the discussion chapter.

4.2 Comparison of prediction results

The figure below shows a comparison of transcription errors from 20 audio samples after transfer learning with Frisian data, highlighting the incorrect outputs generated by the three language models. The audio samples were selected as the first 20 from a randomly shuffled set; therefore, regardless of the random seed used, the content of these 20 samples remained fixed.



Noflik om mei jo yn 'e kunde te kommen.	ft w/ Frisian
English	noflik om mei jo yn e kunde te kommen
Dutch	Noflik om jo yn de kunde te kommen.
German	noflik om mei jo yn e kunde te kommen
regionaal betsjut dat it mar foar in bepaalde krite ornearre is	
English	regionaal betsjut dat it mar foar in bepaalde krite ornearre is
Dutch	regio no betsjut dat it mar foar in bepaalde krite ornearre is
German	regionaal betsjut dat it mar foar in bepaalde krite ornearre is
it kin net fan ien kant komme	
English	it kin net fan de iene kant komme
Dutch	it kin net foar ien kant komme
German	it kin net fan ien kant komme
dat wie de moandeitemiddeis dat wy der wiene net it gefal	
English	dat wie de moandeitemiddeis dat wy der wiene net it gefal
Dutch	dat wie de moandeitemiddeis dat wy der wiene net it gefal
German	dat wie de moandeitemiddeis de twazerien net it gefal
it wikseljend wetterpeil joech swierrichheden by de omwenners	
English	it wikseljende wetterpeil joech swierrichheden by de omwenners
Dutch	it wikseljende wetterpeil joech swierrichheden by de omwenners
German	it wikseljende wetterpeil joech swierrichheden by de omwenners
de frachtwein ferlear in part fan de lading doet de sjauffeur de bocht omgie	
English	de frachtwein ferlear in part fan de lading doet de sjauffeur de bocht omgie
Dutch	de frachtwein ferlear in part fan de lading doet de sjauffeur de bocht omgie
German	de frachtwein fan leger apart fan de lading doet de sjauffeur de bocht omgie
it maklikste is it om op in skriuwblok in tal kolommen te tekenjen	
English	it maklikste is it om op in skriuwblok yn tal kolommen te tekenjen
Dutch	it maklikste is it om op in skriuwblok yn tal kolommen te tekenjen
German	it maklikste is it om op in skriuwblok yn tal kolommen te tekenjen
mei in bytsje gelok treffe jo him yn it park	
English	mei in bytsje gelok treffe jo him yn it park
Dutch	mei in bytsje gelok treffe jo him yn it park
German	mei in bytsje de lulk treffe jo him yn it park
de man remme en bleau oan e kant fan e wei stean	
English	de man remme en bleau oan de kant fan e weistean
Dutch	de man remme en bleau oan de kant fan de wei stean
German	de man remme en bleau him oan e kant fan e wei stean

Figure 2

As shown in the figure above, out of the 20 audio samples, only 9 show differences between the predicted text and the reference transcription. Moreover, not all three language models made errors on these samples, which demonstrates the improvement in model performance after transfer learning.

Next, based on the previously discussed error types, we will analyze the causes of the errors found in these specific samples.

The first sentence,"Noflik om mei jo yn 'e kunde te kommen", contains a transcription error only in the Dutch model. The word "mei" was omitted. This is because "mei" was weakened through liaison in spoken language, making it unrecognizable to the ASR system. This is a case of boundary loss caused by coarticulation or phonetic reduction, which is a typical liaison/phonetic ambiguity error. The word "'e" was transcribed as "de". The first reason is that "'e" and"de" sound similar in pronunciation, and the ASR system selected "de" as it is a more common form in Dutch. Secondly, "'e" is the Frisian form of the definite article "the", while"de" is the common definite article in Dutch. Therefore, this is a language transfer error caused by phonetic similarity, which is a case of homophony.

The second sentence, "regionaal betsjut dat it mar foar in bepaalde krite ornearre is.", contains a transcription error only in the Dutch model. The word "regionaal" was transcribed as"regio no". The original word is an adjective formed from "region" + "-aal", but the ASR system incorrectly split it into "regio" ("region") and "no" (an unknown word). This is a typical compound word error. In addition, the appearance of"no" is likely due to phonetic ambiguity, which caused the system to make an incorrect judgment. This falls under the category of an ambiguous speech region.

The third sentence, "it kin net fan ien kant komme.", contains prediction errors in both the English and Dutch models. First, in the English model, "ien" was transcribed as "de iene". There are two reasons for this. The first reason is that in Frisian,"ien" ("one") and "de iene" ("the one" when used as an adjective with a definite article) are very similar in pronunciation, which makes it difficult for the ASR system to distinguish their actual grammatical function. The model could not tell whether it was being used as a cardinal number or as an adjective with a definite article, so it incorrectly generated "de iene". Since the same speech form has multiple grammatical functions, this phenomenon is a typical case of syncretism. Also, the fact that the ASR model chose the adjective form with a definite article may be due to the model being fine-tuned on Frisian based on the English model, making it more inclined to select the expression "the one", which fits more naturally with English language usage. The Dutch model transcribed "fan" ("from") as "foar" ("for"). "fan" and "foar" can easily be confused in neutral or unstressed syllables, so the ASR system selected the wrong semantic item. This is a homophony error.

The fourth sentence,"dat wie de moandeitemiddeis dat wy der wiene net it gefal.", contains a transcription error in the German model. The German model transcribed "dat wy der wiene" as "de twazerien". The first possible reason is that the original phrase contains multiple short words spoken with liaison, making it difficult for the ASR system to construct recognizable words. This leads to an incorrect result and is classified as an ambiguous speech region error. Secondly, the incorrectly predicted phrase shares some phonetic similarity with the correct one. It is likely that the ASR system misheard parts like "wy der" and"wiene", and then forced them together into a plausible word form, which makes this a case of a compound word error.

The fifth sentence, "it wikseljend wetterpeil joech swierrichheden by de omwenners.", contains the same type of error in the German, English, and Dutch models. All three models transcribed "wikseljend" as "wikseljende". "wikseljend" (neutral/indefinite form) and "wikseljende" (definite/form used before nouns) are different inflected forms of adjectives in Frisian. The fact that all three ASR models replaced the uninflected form "wikseljend" with the inflected "-e" form "wikseljende" suggests that the models tend to generate more frequently used adjective forms. This may be related to the distribution in the training data of Dutch and German, where definite article + adjective structures are common. In the case of English, although this kind of morphological variation does not exist, the model might have picked up the inflected form due to its exposure to Frisian data during training, even though such a word form does not exist in English. This phenomenon is a typical case of syncretism.

The sixth sentence, “de frachtwein ferlear in part fan de lading doet de sjauffeur de bocht omgie.”, contains a transcription error only in the German model. The German model transcribed “ferlear in part” as “fan leger apart”. In the original sentence, “ferlear in part” is spoken with liaison at natural speed, causing the ASR model to incorrectly split the phrase into “fan leger apart”. Here, “fan” is a common Frisian preposition meaning “from”, “leger” is a word in German meaning “casual”, and “apart” is a frequent word. Therefore, the model’s segmentation into these three words is understandable, and this error is classified as a phonetic ambiguity (ambiguous speech region) error. This error appears only in the German model, possibly because the German model is more sensitive to phoneme blending, while the Dutch and English models are relatively more conservative, and thus did not produce this mistake.

The seventh sentence, “it maklikste is it om op in skriuwblok in tal kolommen te tekenjen.”, contains the same error across all three models, where “in” was transcribed as “yn”. This is a typical case of syncretism: in Frisian, the indefinite article “in” and the preposition “yn” are very similar in pronunciation when spoken with reduction or liaison (/ɪn/ vs /ən/) . Because of this, the ASR system has difficulty distinguishing their grammatical functions, and therefore mistakenly recognized the indefinite article “in” as the preposition “yn”.

The eighth sentence, “mei in bytsje gelok treffe jo him yn it park.”, contains a prediction error only in the German model. The model transcribed “gelok” as “de lulk”. In Frisian, “bytsje gelok” is a common expression, but it contains multiple weakly stressed syllables. When the ASR system encounters segments with unclear syllable boundaries, it tends to fill in the gaps using more frequent words from its language model, which in this case resulted in “de lulk”. This error is classified as an ambiguous speech region error.

The final sentence, “de man remme en bleau oan ’e kant fan ’e wei stean.”, contains transcription errors in all three models, and each model made different types of mistakes. First, in the English model, “’e” was transcribed as “de”. This is because “’e” is the contracted definite article “the” in Frisian, pronounced in a very reduced form, usually as /ə/ . The English ASR model mistakenly recognized it as the more common and clearer form “de”, making this a typical homophony error. Additionally, it transcribed “wei stean” as “weistean”, likely because in spoken language, this phrase is frequently connected by liaison, causing the ASR system to incorrectly merge the two into a new, incorrect word. This is an example of an ambiguous speech region error. Second, in the Dutch model, all instances of “’e” were also transcribed as “de”, which is the same type of homophony error found in the English model. Finally, the German model transcribed “bleau oan” as “bleau him oan”, possibly because “bleau oan” was spoken with unclear liaison, making the word boundaries hard to detect. The ASR system filled in the ambiguous acoustic space with a higher-frequency word, resulting in “him” being inserted. This is another case of a phonetic ambiguity error.

A summary of these patterns and a response to the research hypotheses will be presented in the next chapter.

5 Discussion

This chapter provides an explanation of the experimental results, focusing on the reasons why the English-based model achieved the best WER performance after transfer learning compared to the other language models. It also offers a summary of the identified error types observed in the predictions. In addition, this chapter discusses the limitations of the experiment and highlights the contributions of the study to the field of cross-lingual speech recognition.

5.1 Validation of Hypothesis

In this subsection, we aim to verify the experimental hypothesis. The first hypothesis states: "Fine-tuning the Dutch, German, and English-based zero-shot Whisper-small models on Frisian data will lead to a significant improvement in ASR performance." Additionally, the study assumed that the Dutch model would perform the best. Firstly, the experimental results do confirm that after fine-tuning on Frisian data, all three language models showed substantial performance improvements. However, the English model ultimately achieved the best results, which does not align with our original hypothesis.

Therefore, the underlying reasons for the results can be interpreted as follows: First, the performance of the Whisper-small model is likely influenced by the size of the English training data. As a multilingual model, Whisper was trained on corpora of varying sizes across different languages. English, being a high-resource language, constitutes a large portion of the training data. In contrast, Dutch and German make up a smaller share. Although both languages are linguistically related to Frisian, their limited representation in the training data may constrain their adaptability during fine-tuning. As a result, the English-based model demonstrated superior recognition accuracy after fine-tuning.

In cross-lingual transfer learning, the degree of linguistic similarity between languages plays a crucial role in model performance. Although Dutch and German are geographically close to Frisian, Frisian shares a closer historical and linguistic relationship with English. Both Frisian and Old English belong to the same branch of the West Germanic language family and exhibit notable similarities in pronunciation, vocabulary, and syntactic structures(Van Heuven & Kirsner, 2004). In contrast, while Dutch and German share some lexical items with Frisian, they differ more substantially in sentence structure, grammatical inflection, and phonological phenomena such as liaison. These differences may increase the likelihood of misinterpretation during transfer. Therefore, the superior performance of the English-based model after fine-tuning is likely attributable not only to the larger amount of training data but also to its closer linguistic proximity to Frisian.

So two factors likely explain English's superior performance: the larger training corpus typically available for English models and the closer historical linguistic relationship between English and Frisian(Van Heuven & Kirsner, 2004).

As for the second hypothesis:"The differences in model performance across source languages will be influenced by linguistic factors, such as the lexical and syntactic differences between the source language and Frisian.And the richness and differences in language structure may lead to Compound

Word Errors." Based on the comparison of experimental results and the analysis of errors using the six linguistic causes, the answer is definitely yes.

More specifically, this conclusion can be drawn from the detailed explanations of each transcription error given above. In summary, it can be seen that most of the errors fall into the categories identified in existing linguistic research, particularly phonetic ambiguity, syncretism, and homophony. Many of these errors occur in fast or connected speech, where it becomes difficult for the ASR system to accurately detect word boundaries—for example, "ferlear in part" being misheard as "fan leger apart", and "wei stean" being merged into "weistean". Confusions such as "in" vs. "yn" and "ien" vs. "de iene" reflect syncretism, where similar-sounding forms have different grammatical functions. The frequent replacement of the definite article "'e" with "de" shows a cross-linguistic interference pattern caused by similar phonetic forms. This also indicates that the ASR model tends to prefer more frequent word forms from its training data, leading to homophony errors.

Overall, the German model was more prone to reconstructing unclear speech into incorrect new words, the English model occasionally produced invented words in liaison-heavy environments, while the Dutch model showed a stronger tendency toward grammatical "standardization". Therefore, we can conclude that ASR systems are heavily influenced by their internal language preferences, especially in a multilingual model setting.

5.2 Limitations

5.2.1 The available data is not large

The current study relies on a single source of Frisian data—Common Voice Corpus 21.0, which provides 70 hours of audio. While useful, this dataset is relatively small in size and limited in scope, as it only includes West Frisian and lacks coverage of other Frisian language branches such as East Frisian or North Frisian. Consequently, the linguistic diversity of the data is limited, which may restrict the generalizability of the findings.

5.2.2 The model has a limited size

Due to constraints in research time and hardware resources, this study only used the Whisper-small version. Although this version remains the generalization capability of the original model, it also has several limitations. For example, while Whisper-small supports multiple languages, its smaller parameter size means that its cross-lingual knowledge representation is more limited. In cases involving phonetic ambiguity or liaison, the model's capacity may be insufficient to distinguish unclear boundaries, leading to a higher rate of errors—an issue clearly reflected in the error analysis presented in the previous chapter.

Moreover, in the architecture of the Whisper-small model, the number of Transformer decoder layers and attention heads is reduced. This limits the model's ability to capture long-range contextual dependencies, which can result in mistakes of longer sentences.

5.2.3 The evaluation metric is limited

The evaluation metric used in this study was limited to WER. Although WER is one of the most common metrics, relying solely on this single criterion presents certain limitations. WER calculates errors—insertions, deletions, and substitutions—at the word level. In contrast, using more fine-grained metrics such as PER (Phoneme Error Rate) or CER (Character Error Rate) could assess the model's recognition accuracy at the phonemic or morphological level, leading to a more detailed evaluation.

5.3 Future work

Future research can continue along the following two directions:

First Research Question: How can the trade-off between performance and computational cost be managed when using different sizes of Whisper models for low-resource ASR tasks? It is recommended to systematically compare the relationship between model size and performance, focusing on the following three aspects:

(1) The improvement in Word Error Rate (WER) per additional parameter;

(2) The computational cost of training and inference (including time and memory usage);

(3) Changes in finer-grained evaluation metrics such as Character Error Rate (CER) across different model sizes.

These comparisons can help identify the most suitable model size for real-world settings where computational resources are limited.

Second Research Question: How do different Frisian dialects perform in transfer learning scenarios?

This study used only West Frisian. Future work could extend to North Frisian and East Frisian. By conducting cross-dialectal comparisons, researchers can not only further verify the generalizability of models across Frisian language variants but also explore the impact of linguistic distance on transfer learning performance.

6 Conclusion

This study successfully verified the feasibility of applying transfer learning to the low-resource language Frisian using the Whisper-small model and three related high-resource languages: Dutch, German, and English. The ASR model performance improved significantly. It also provided a linguistic analysis of the errors after transfer learning and found that most errors were due to phonetic ambiguity, syncretism, and homophony. Furthermore, it showed that the ASR model's behavior in multilingual settings is influenced by the internal characteristics of its source language.

Although the study achieved promising results, it also had some limitations, such as using a single data source, a small model size, and relying on a single evaluation metric. Future work will focus on expanding the dataset, testing larger models, and adopting more diverse evaluation methods.

The achieved WER reduction from $\sim 100\%$ to $\sim 6\%$ represents more than technical progress—it directly addresses the ethical imperative of Frisian language preservation discussed in Section 3.4. By demonstrating that high-quality ASR can be developed through transfer learning without extensive Frisian-specific resources, this work provides a cost-effective pathway for creating practical language technologies that support Frisian speakers in daily life. Such technologies are crucial for language vitality: they enable digital accessibility for elderly speakers, facilitate educational tools for younger generations learning Frisian, and ensure the language remains relevant in modern digital contexts. The linguistic error analysis further contributes by identifying specific challenges (liaison, syncretism) that developers must address to create truly inclusive ASR systems that work across Frisian dialects and speaker demographics. This approach thus fulfills our ethical commitment to supporting linguistic diversity through accessible technology development.

6.1 ASR applications for Frisian communities

This study holds significant practical value and social relevance for the Frisian-speaking community. By leveraging the Whisper multilingual pre-trained model in combination with cross-lingual transfer learning strategies, we achieved substantial improvements in speech recognition performance using only a limited amount of West Frisian data—without the need to develop dedicated phoneme lexicons, language models, or complex preprocessing pipelines.

This low-barrier, reproducible approach offers a viable technological pathway for local educational institutions, language technology developers, and cultural organizations. It has the potential to accelerate the adoption of Frisian in applications such as voice input, automatic subtitling, and speech retrieval. At the same time, it provides new momentum for the digital preservation and revitalization of the Frisian language, enhancing its visibility in media, public services, and language education.

Although this study focuses on West Frisian, the proposed strategy can be extended to other Frisian dialects to support cross-dialect recognition. Overall, the study offers a promising framework for minority language communities to develop speech recognition systems under low-resource conditions, contributing to the advancement of language technology.

In conclusion, this study fills a gap in transfer learning between related languages and offers a linguistic perspective on ASR errors, contributing to the advancement of speech recognition technologies for low-resource languages.

References

Adda-Decker, M., Vasilescu, I., Snoeren, N., Yahia, D., & Lamel, L. (2011). Towards exploring linguistic variation in asr errors: paradigm and tool for perceptual experiments. *New Tools and Methods for very-large-scale phonetics research*.

Amooie, R., de Vries, W., Hao, Y., Dijkstra, J., Coler, M., & Wieling, M. (2025). Evaluating standard and dialectal frisian asr: Multilingual fine-tuning and language identification for improved low-resource performance. *arXiv preprint arXiv:2502.04883*.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., … Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., … others (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Bălan, D. A. (2023). *Improving the state-of-the-art frisian asr by fine-tuning large-scale cross-lingual pre-trained models* (Unpublished doctoral dissertation).

Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2018). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, *56*, 85–100.

Choe, J., Chen, Y., Chan, M. P. Y., Li, A., Gao, X., & Holliday, N. (2022). Language-specific effects on automatic speech recognition errors for world englishes. In *Proceedings of the 29th international conference on computational linguistics* (pp. 7177–7186).

de Graaf, T. (2016). Dutch, frisian and low german: the state language of the netherlands and its relationship with two germanic minority languages. part 1.

de Zuazo, X., Navas, E., Saratxaga, I., & Rioja, I. H. (2025). Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*.

Dyk, T. d., Siebren; Vries. (undefined). *Personal pronouns.* Retrieved from `https://taalportaal.org/taalportaal/topic/pid/topic-13998813311277191` (Retrieved June 10, 2025 from https://taalportaal.org/taalportaal/topic/pid/topic-13998813311277191)

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

Jong, E., Gerbrich de; Hoekstra. (2015, December). *A general introduction to Frisian.* Retrieved from `https://taalportaal.org/taalportaal/topic/pid/topic-14225224491227143` (Retrieved June 10, 2025 from https://taalportaal.org/taalportaal/topic/pid/topic-14225224491227143)

Khurana, S., Laurent, A., & Glass, J. (2022). Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6647–6651).

Kozhirbayev, Z. (2023). Kazakh speech recognition: Wav2vec2. 0 vs. whisper. *Journal of Advances in Information Technology*, *14*(6), 1382–1389.

Kumar, R., Adiga, D., Ranjan, R., Krishna, A., Ramakrishnan, G., Goyal, P., & Jyothi, P. (2022). Linguistically informed post-processing for asr error correction in sanskrit. In *Interspeech* (pp. 2293–2297).

Li, B., Chang, S.-y., Sainath, T. N., Pang, R., He, Y., Strohman, T., & Wu, Y. (2020). Towards fast and accurate streaming end-to-end asr. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6069–6073).

Liu, W., Fu, K., Tian, X., Shi, S., Li, W., Ma, Z., & Lee, T. (2023). An asr-free fluency scoring approach with self-supervised learning. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Pujol, P., Pol, S., Nadeu, C., Hagen, A., & Bourlard, H. (2005). Comparison and combination of features in a hybrid hmm/mlp and a hmm/gmm speech recognition system. *IEEE Transactions on Speech and Audio processing*, *13*(1), 14–22.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Ringe, D., & Taylor, A. (2014). *The development of old english* (Vol. 2). OUP Oxford.

Tu, T., Chen, Y.-J., Yeh, C.-c., & Lee, H.-Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*.

Van Heuven, V. J., & Kirsner, R. S. (2004). Phonetic or phonological contrasts in dutch boundary tones? *Linguistics in the Netherlands*, *21*(1), 102–113.

Visser, W. (2015, December). *Breaking.* Retrieved from `https://taalportaal.org/taalportaal/topic/pid/topic-14358346024113628` (Retrieved June 10, 2025 from https://taalportaal.org/taalportaal/topic/pid/topic-14358346024113628)

Wang, C., Pino, J., & Gu, J. (2020). Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation. *arXiv preprint arXiv:2006.05474*.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent variable analysis and signal separation: 12th international conference, lva/ica 2015, liberec, czech republic, august 25-28, 2015, proceedings 12* (pp. 91–99).

Winter, C. (2022). Frisian. In *Oxford research encyclopedia of linguistics.*

Yadav, H., & Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2202.12576*.

Żelasko, P., Moro-Velázquez, L., Hasegawa-Johnson, M., Scharenborg, O., & Dehak, N. (2020). That sounds familiar: an analysis of phonetic representations transfer across languages. *arXiv preprint arXiv:2005.08118*.

Appendices

A Declaration of AI use

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used ChatGPT 4o and DeepL translation for the following purposes:

Content generation :

– Paraphrasing cited content in Chapter 2: I used ChatGPT 4o to rewrite and clarify complex source texts. I selected and interpreted the citations myself; AI was used to generate clearer paraphrasing in English without altering meaning.

– Alternative explanations for technical concepts (Chapter 3, Section 3.1): I used ChatGPT 4o to generate clearer formulations of complex technical material. The underlying ideas and sources were selected and verified independently by me.

– Literature summarization (preliminary stage): I used ChatGPT 4o to summarize some background readings during early stages of the project. These summaries were used only for orientation and were not used verbatim in the final thesis.

– Grammatical structure support in ASR error analysis (Chapter 4, Section 4.2): I used ChatGPT 4o to assist in identifying grammatical structures in selected sentences related to ASR errors. The interpretation and integration into the error analysis were conducted independently by me.

And I also used AI tools for creating initial code documentation templates. All AI-generated content was reviewed, verified, and edited by me to ensure accuracy, appropriateness, and alignment with academic standards.

Xinchi Li / 10.06.2025