university of groningen

campus fryslân

Addressing ASR Bias Against Foreign-Accented Dutch: A Synthetic Data Approach

Maria Tepei

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Addressing ASR Bias Against Foreign-Accented Dutch: A Synthetic Data Approach

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
Phat Do, MA (Voice Technology, University of Groningen)

Maria Tepei (S5713544)

June 11, 2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis supervisor, Phat Do. Not only were his guidance and technical support invaluable throughout the process of completing my thesis project, but the knowledge and practical skills I have acquired in the classes taught by him (Speech Recognition I and Speech synthesis II) have been instrumental in shaping this study; I am profoundly grateful for his mentorship.

I also extend my sincere thanks to Dr. Matt Coler, the program director, whose years of effort have made this study program possible, for his professional support, as well as the constant encouragement throughout my academic and professional growth. At the same time, I am indebted to all other lecturers in the MSc Voice Technology program (Dr. Joshua Schäuble, Dr. Vass Verkhodanova, and Dr. Shekhar Nayak), as their dedication and expertise have brought me from "Hello, world!" to training and fine-tuning state-of-the-art speech recognition and speech synthesis models in less than a year's time.

I acknowledge the Centre for Information Technology at the University of Groningen for providing access to and support for the Habrok high-performance computing cluster; the extensive computations required for my final thesis project would have not been possible without such resources.

I would also like to acknowledge the immense gratitude I have for my parents, Delia and Gheorghe. They have made countless sacrifices since the very beginning of my education years, driven by their strong belief in the power of knowledge, and studying abroad would have not been possible without them.

Last, but not least, I thank my partner, Alex, who has been my immediate support through one the most stressful times of my life, showing me endless patience and loving care. Additionally, his statistical expertise has played an important role in double-checking the soundness and accuracy of my statistical analysis.

Abstract

Despite substantial improvements in automatic speech recognition (ASR) over the last years, the high performance achieved for ”standard speakers” does not hold across all genders, ages, or foreign accents. As a result, an important area of research is inclusive ASR, aimed at reducing the performance gaps such systems display across subgroups of the population. In the present thesis, I evaluate one of the most recent and robust ASR systems (OpenAI's Whisper) to uncover and assess the level of bias it displays against foreign-accented Dutch. Additionally, I investigate whether synthetically accented speech samples obtained from a fine-tuned speech synthesis model (FastSpeech2) can act as a viable data augmentation tool to create additional training data for Whisper, in a fine-tuning transfer learning paradigm. By investigating bias, as opposed to WER reduction, I specifically pay attention to both the improvement in performance on foreign-accented Dutch and the potential decrease in performance on native Dutch. Experimental results show that fine-tuning Whisper on synthetic accented speech data does increase its performance on natural speech samples, although this comes at the cost of decreased performance on native samples after fine-tuning. Additionally, the insights from fine-tuning Whisper put into question its suitability for this learning paradigm, as its large number of parameters displays increased stability on small, low-resource datasets.

Keywords: bias mitigation; accented Dutch speech; data augmentation; voice conversion; fine-tuning Whisper

Contents

1 Introduction

Automatic speech recognition (ASR) as an area of research has made remarkable progress over the past decades, achieving performance on par with, if not surpassing, the human accuracy on speech transcription tasks. These advancements have been driven by the development of sophisticated models and algorithms tasked with transcribing speech, which are at this point performing almost perfectly – as long as the speaker is a white, highly-educated young woman with a US English accent and no speech impairment (Feng, Halpern, Kudina, & Scharenborg, 2024; Feng, Kudina, Halpern, & Scharenborg, 2021; Fuckner, Horsman, Wiggers, & Janssen, 2023; Martin & Wright, 2023; Zhang, Herygers, Patel, Yue, & Scharenborg, 2023). In other words, despite impressive technological leaps, numerous studies indicate that ASR systems still face significant challenge when dealing with speech that is underrepresented in, if not absent from the training data, highlighting the persistent challenge of achieving equitable performance across diverse speech profiles.

Foreign-accented has been identified as one of the primary sources of bias, second only to gender (Benzeghiba et al., 2007), and its complex nature makes it a persistent challenge in the field of inclusive ASR to this day (Feng et al., 2024). This is primarily because accented speech involves fine-grained modifications in the acoustics of speech, which stem from altered word pronunciations, coarticulations, substitutions, reductions, or non-standard pitch trajectories and speech rate. Moreover, the nature of these acoustic shifts is highly dependent on the speaker's native language (L1), which means that handling different accents by a single ASR system is an added challenge. Additionally, simply incorporating a subset of accented speech samples in the batch of training data is not enough to successfully mitigate accent bias, for two main reasons. The intricate and irregular acoustic variation injected by foreign accent into the speech characteristics, as well as morphosyntactic shifts such as ungrammatical word order, vary from one accent to another, as well as from one speaker to another, which leaves little regularity for the model to learn during training. Moreover, especially in scenarios with very limited accented data (which is rather the norm than an exception when it comes to foreign-accented speech), ASR models often fail to disentangle accent-specific acoustic features from speaker identity, thus being prone to overfit on speakers present in the training data, yet perform very poorly on unseen speakers with the same accent.

1.1 Motivation

With increased mobility across countries becoming a global trend, several Western languages such as English, Spanish, French, or Dutch have an increasing population of non-native speakers; for example, more than three million people living in The Netherlands have a non-Dutch background. This means that more of our attention, research efforts, and resources need to go towards developing ASR systems that do not perform differently across speaker subgroups. Specifically, not only do we need to find a way for ASR systems to perform well on accented speech, but more importantly we need to find ways in which the performance gap across speaker groups (i.e. the bias) closes.

Previous attempts at addressing bias against foreign-accented speech have focused on either improving the training data (e.g. via data augmentation) or changing the ASR architecture and training strategy. The main limitation of architecture-driven approaches is that both excellent and poor performances are often hard to interpret and explain directly; on the other hand, data-driven approaches are prone to running into the data scarcity issue, with accented speech samples often being very limited, if at all existent. The present study builds on previous approaches and extends them in several ways.

One of the primary motivations for my research is to address potential limitations of the dataset

that has been used for bias mitigation on accented Dutch so far. Previous experiments in the same line (Zhang, Zhang, Halpern, Patel, & Scharenborg, 2022; Zhang, Zhang, Patel, & Scharenborg, 2022) predominantly use the JASMIN-CGN corpus (Cucchiarini, Driesen, Hamme, & Sanders, 2008) to generate synthetic accented speech via voice conversion, which provides a good standard for comparability purposes, but may not be optimal for speech synthesis tasks. The JASMIN-CGN corpus was design for ASR tasks and thus comprises of recordings that contain noise, volume and speech rate variations, or inconsistent pitch contours; while these are excellent factors for testing ASR robustness, they can hinder the quality of the result when used for speech synthesis or voice conversion, as the latter work best on clear, high-quality, consistent data. For this reason, my study experiments with a custom dataset of high-quality read speech, which I have collected specifically with the speech synthesis task in mind. At the same time, I hope to address the replicability issue by making the recordings public and available for further use.

Another motivation stems from taking into account the latest ASR architectures. Previous studies investigating accent bias in Dutch speech have made use of older ASR models, such as an RNN-based or transformer-based sequence-to-sequence architecture in (Zhang et al., 2023) or a TDNN-LSTM architecture in Feng et al. (2021) and Zhang, Zhang, Patel, and Scharenborg (2022). The whisper model, however, has since become the state of the art in ASR, displaying low WERs across several languages, out of which the lowest WER on the Common Voice dataset is for Dutch. This makes whisper a promising ASR architecture for investigating robustness to foreign accent, which I investigate in my experiments.

Lastly, I explore a different approach to data augmentation by using a speech synthesis model. Traditional data augmentation techniques such as speed perturbation, pitch shifting, noise addition, or SpecAugment can be effective, but often fall short in addressing qualitative variations in accented speech. Recent experiments (Klumpp et al., 2023; Zhang et al., 2023) show the potential of voice conversion techniques to improve ASR performance in low resource contexts by generating supplementary data which improves the training set qualitatively, as well as quantitatively. Unfortunately, high-performing voice conversion models which work in zero-shot settings (Jin et al., 2023; Quamer, Das, Levis, Chukharev-Hudilainen, & Gutierrez-Osuna, 2022) do not have openly available code that can enable replication, while openly-available models such as AGAIN-VC (Chen, Wu, Wu, & Lee, 2020) cannot disentangle accent features from speaker identity given a single speaker dataset. Furthermore, many such models require parallel training data, specifically utterance pairs of native and foreign-accented speech samples with the same linguistic content – a difficult requirement for datasets of low-resource nature such as foreign-accented speech. As a result, I approach the task of creating artificial accented data by means of a speech synthesis model, FastSpeech2, by training it on native Dutch and subsequently fine-tuning it on a small, single-speaker dataset to obtain high-quality, synthetic accented Dutch speech.

1.2 Research Questions and Hypotheses

I adopt a data-driven approach to addressing bias against non-native Dutch speech by exploring pre-existing architectures (FastSpeech2 for speech synthesis and Whisper for speech recognition) from an evaluative perspective. The main research question I aim to answer is:

> **RQ**1: Does synthetic accented Dutch speech from a fine-tuned FastSpeech2 model improve the performance of Whisper on natural foreign-accented speech?

> **H1**: I hypothesise that fine-tuning the Whisper model on synthetic accented Dutch speech will improve its performance on natural accented Dutch speech. In other words, I expect the WER of the fine-tuned model on accented speech to be lower than the WER of the pre-trained, out-of-the-box Whisper on the same accented dataset.

Hypothesis 1 is in line with previous studies which have shown that improved performance across various ASR architectures can be obtained by using various data augmentation techniques. For example, Klumpp et al. (2023) show in their experiments on English and various accent flavours that synthetically accented speech samples, obtained by means of a voice conversion model, improve an ASR model's robustness to foreign-accented pronunciation alternatives, although the authors note that this did not bring about a *cross-accent* robustness, meaning that the model was not performing better on unseen accents as well. In a similar vein, Zhang, Zhang, Halpern, et al. (2022) investigate the case of accented Dutch speech specifically and the viability of several data augmentation techniques in closing the performance gap, finding that speed perturbation in combination with synthetically-accented speech samples from another voice conversion model (Chen et al., 2020) yielded the best results in bias reduction. While I use FastSpeech2, which is a speech synthesis model and not a voice conversion model, the ultimate aim is the same: to create foreign-accented speech samples by synthetic means, from a limited amount of natural accented data.

In order to better situate my findings in the field and to strengthen the motivation for my study, I aim to also answer two related questions. Firstly, RQ1 contains the underlying assumption that without fine-tuning, Whisper will perform significantly worse on natural accented speech compared to native speech. I will confirm whether this is the case by looking into the following:

> **RQ2**: Does out-of-the-box Whisper perform significantly worse on accented Dutch speech compared to native Dutch speech?

> **H2**: I hypothesise that Whisper will still display the persistent performance gap across accented and native speech, as documented in previous literature. In statistical terms, I expect a significantly higher WER on the natural accented data compared to the WER on the native data before fine-tuning.

I expect Whisper to display the gap in performance between native and foreign-accented Dutch speech that has been widely documented in previous literature across various ASR architectures. It is true that Whisper is a recent model that has showed outstanding performance across tasks and languages, as reviewed below; however, several review studies (Benzeghiba et al., 2007; Zhang, Zhang, Halpern, et al., 2022) point out the persistent challenge of foreign-accented speech and its associated feature shifts, which is likely impossible to completely overcome even by the most recent ASR systems.

Lastly, I want to investigate whether this approach truly closes the performance gap instead of simply shifting it; more specifically, not only is it important for the fine-tuned model to perform better (i.e. have a lower WER) on accented Dutch, but it should also maintain its initial performance on native speech to a similar level. To this end, the third research question in my study is:

> **RQ3**: Does the model's performance on native Dutch speech significantly degrade after the fine-tuning process?

> **H3**: In line with theoretical aspects of fine-tuning as a learning strategy, I expect the fine-tuned model's performance to slightly go down on native Dutch speech, compared to its pre-trained counterpart. Whether this degradation is significant or not remains to be confirmed by the corresponding statistical test.

Catastrophic forgetting (Wang & Chen, 2023) is a known phenomenon that can occur when a model pre-trained on one task (here: cross-language ASR and translation) and fine-tuned on a downstream, more specific task (here: transcribing foreign-accented Dutch from a single speaker) exhibits a decrease in performance on the original task. During fine-tuning, Whisper undergoes backpropagation to minimise the loss on the new dataset, updating its parameters; consequently, weights are significantly altered to fit the new data, which often degrades performance on the previously learned task. Although I try to mitigate this by implementing parameter-efficient fine-tuning, I still expect it to happen to some degree.

By offering data-driven answers to these questions, I hope to further explore how promising speech-synthesis-driven data augmentation truly is in addressing ASR bias against non-native Dutch speech, particularly given its most common challenge: data scarcity.

1.3 Outline

The sections of this thesis are organised as follows. Section 2 presents a review of previous literature in the field, the main challenges associated with accented speech recognition, the primary sources of bias, as well as a summary of previous attempts at mitigating this bias. Section 3 details the methodological approach adopted in the present study, including descriptions of the used datasets, the models, and the experimental conditions under which they are tested. Section 4 reports the results and Section 5 discusses the in detail, establishing the answers to the research questions and situating them in the field. Finally, Section 6 concludes the thesis by summarising the main findings, their relevance in the field of inclusive ASR, as well as the directions for future research that it opens. Further details and supplementary materials can be found in the Appendices.

2 Literature Review

2.1 Challenges in Accented Speech Recognition

Despite remarkable recent advancements in the field of automatic speech recognition, the inherent variability of human speech continues to pose challenges for such systems, due to factors such as gender, age, or foreign accent affecting the fine acoustics of each individual's speech. In a systematic review of sources of speech variability, Benzeghiba et al. (2007) identify gender and foreign accent as the two primary sources of variability affecting ASR performance.

While gender-related variation is relatively easy to address through balanced datasets, accent-related variations pose a more intricate challenge due to their complex nature. Accent alters the fine acoustic structure of speech (e.g. voice quality), but it also manifests as modified word pronunciations, yielding coarticulations, substitutions, reductions, or non-standard pitch trajectories. Moreover, the *type* of acoustic shifts introduced by foreign accent depends on the native language of the speaker (L1), while the *perceptual salience* of those shifts depend on their level of proficiency in the target non-native language (L2).

Feng et al. (2021) use a data-driven approach to investigate bias in a DNN-based ASR system for Dutch speech, where both the objective WER measure and a subsequent qualitative analysis show that bias stems from various factors, of which the main one is the composition of the training dataset, which reflects the level of variations in speaking styles and accents, or differences related to vocal tract characteristics, such as age or gender. The authors use speech data from the Corpus of Spoken Dutch – CGN (Schuurman, Schouppe, Hoekstra, & Van der Wouden, 2003) and its more recent JASMIN-CGN extension (Cucchiarini et al., 2008). The results indicate a bias against male speech, which is less accurately recognised in general, but also in correlation with accentedness, i.e. non-native female speech is more accurately recognised than native male speech. Interestingly, when it comes to bias against non-native speech (across genders and ages), the results show that an increased L2 proficiency does not correlate with a WER reduction for the L2-speakers of Dutch, which the authors explain in terms of the nature of CEF evaluations of proficiency: lower CEF-levels focus more on vocabulary and grammar rather than pronunciation, which means that CEF-based proficiency level cannot act as a reliable proxy for accent strength. At the same time, a qualitative analysis of phoneme error rate indicates that vowels and diphthongs which are notoriously difficult to acquire by L2 learners of Dutch are also the most challenging sounds for the ASR system to recognise and transcribe accurately.

The impact of foreign-accented speech on the performance of ASR systems is a relatively small, but increasingly active area of research within the field of speech technology, especially as the focus in the field shifts from performance-driven research goals towards extending the results and generalise the high-performance architectures across all speaker groups. The impact of foreign-accented speech on ASR systems includes a shift in the acoustic feature space (Benzeghiba et al., 2007), which is contingent upon the speaker's native language (L1), but also correlates with factors such as gender or age (Feng et al., 2021). An intuitive solution to this would be to simply include dialectal or accented data into the training sets, in order to familiarise the model with alternative pronunciations and thus increase its robustness to accented pronunciations. However, such approaches fall short, as the foreign-accent-induced acoustic and morpho-syntactic shifts are rather intricate and not always regular, which leaves limited patterns for the system to exploit during training. Consequently, the incorporation of accented data during training has to be rather generous (which is often impossible due to the scarcity of accented speech data), otherwise it will get lost among the better-represented native samples and thus fail to significantly improve the model's performance. This underscores the

need for more sophisticated approaches to mitigate accent-related variability.

Overall, previous literature points to multiple challenges in ASR for foreign-accented speech which persist to this day in this field. One major challenge is data scarcity, coupled with the variety of L1-L2 possible combinations; it is simply expensive and often not feasible to have even small datasets of every language spoken with every different foreign accent. Moreover, merely having speech data of accented speech is sometimes not enough, because dialects or feature shifts often do not happen at a strictly acoustic level, but also at lexical and morpho-syntactic levels (e.g. non-standard word order, regional words), which means that other resources used, such as the pronunciation dictionary or the language model, need to be gathered, trained, or fine-tuned as well. Efforts to overcome these challenges focus usually either on the model architecture (i.e. 'making the most of what you have', by improving or adapting the architecture for a low resource or zero resource setting) or on the data used (various kinds of data augmentation), if not a combination of both.

2.2 Addressing bias against non-native speech - Previous approaches

2.2.1 What is bias?

Bias is commonly defined in the literature as a gap in the performance of an ASR system across groups, usually quantified in terms of word error rate (WER), phoneme error rate (PER), or – more finely grained – character error rate (CER) (Feng et al., 2024, 2021). Martin and Wright (2023) more specifically adopt the definition of bias as 'cases where computer-based systems systematically and unfairly discriminate against individuals or groups of individuals in favour of others' (p. 617). Following this definition, a biased ASR system is one that systematically functions more poorly for the speech of a subgroup, which results in unequal and disadvantageous outcomes when people belonging to these subgroups interact with the systems.

Moreover, to elucidate the sources of bias, the authors distinguish several possible *loci* along the development pipeline where bias could seep in. Pre-existing bias is related to the composition of the development team, comprising of the values and attitudes of its members, as well as the vision of the funding institution (Kudina, 2024). These are reflected in the subsequent development strategies and decisions, creeping into the training data and ultimately mirroring the biases of the society from which the final product originates. Bias can also emerge post-development (emergent bias), if a system specifically designed for one context is applied to a different one (e.g. an ASR designed to transcribe healthy speech is employed in transcribing the speech of people with dysarthria). Overall, heavily biased ASR systems can yield negative outcomes through allocation or representational harms, if a certain subgroup is for instance misrepresented, or their existence is not recognised or acknowledged altogether (Martin & Wright, 2023).

When it comes to sources of bias in recognising and transcribing Dutch speech, Feng et al. (2021) identify significant biases related to gender, age, and foreign accent. Specifically, male voices are recognised less accurately compared to female voices, while across the age axis, teenage speech is recognised most effectively, followed by elderly speech and lastly by child speech, which is the least accurately recognised. Additionally, the authors reveal that ASR systems exhibit a notable bias against non-native and regional varieties of Dutch. A similar observation is made by Fuckner et al. (2023), who compare the performance of two state-of-the-art ASR systems on accented Dutch speech using the same dataset and find that OpenAI's whisper-small model largely outperforms Meta's Wav2Vec2 on accented data across genders and ages.

A more detailed phoneme analysis, employed by both Feng et al. (2021) and Fuckner et al. (2023)

suggested that differences in pronunciation, particularly regarding phonemes which are known to be hard to acquire by L2 speakers of Dutch (/œy/, /Y/, /y/, and /ø:/), have the biggest contribution to the disparity in recognition accuracy across native and non-native groups. Thus a critical source of bias is found in the training material itself: while the overall speaker distribution in the training data did not directly account for the observed biases, a subsequent analysis indicated that phonemes which are underrepresented and difficult to acquire are frequently and systematically misrecognised. Furthermore, biases related to the ASR architectures used also exist, though these are more complex and less transparent, necessitating further investigation to fully understand their implications.

While such theoretical observations and taxonomies might initially seem rather abstract, it is crucial to recognise their real, practical implications as (voice-based) AI systems are increasingly integrated into various fields and their decision-making processes are more heavily relied upon across industries. Such biases can significantly impact the lives of end users, for instance in voice-based examinations or pronunciation training, where ASR performance may disadvantage speakers whose accents deviate from standardised training data or their speech might be wrongfully flagged as incorrect or ungrammatical. Furthermore, biases of voice-based chatbots, increasingly adopted in customer relations, healthcare, human resources departments, or as a point of contact with public institutions, can lead to unequal access to private or public services. As ASR systems are implemented across industries and employed more and more by non-experts, it is imperative to first understand and subsequently address these biases, thus aiming for fair and equitable outcomes for all users.

At the same time, it is important to note that bias in ASR systems is inevitable (Kudina, 2024). Speech data intrinsically carries demographic biases, reflecting latent information about participants' age, gender, and ethnic background. Therefore, the ultimate goal of bias mitigation in ASR is not to entirely eliminate bias, as that is not achievable. Instead, the desirable outcome is to remain aware of bias, identify its sources and influencing factors, understand its implications for the final product, and increase the system's robustness to expected feature shifts in the speech of all end users across social groups. Solutions such as data-driven, explainable approaches to bias mitigation can help developers understand how and to what extent the large corpora used in training acoustic and language models are representative of various user groups. This understanding ensures the final model's robustness to the associated variability in input speech.

Attempts at reducing bias against non-native and dialectal varieties commonly fall in one of two categories. Architecture-driven approaches make use of the latest technical advancements to address the model's robustness to pronunciation variations, especially given the limited or sometimes absent accented speech data. These methods, however, are prone to functioning as 'black boxes', producing results and making decisions and generalisations which are difficult to interpret. On the other hand, data-driven approaches can be integrated with existing architectures, albeit sometimes slightly tweaked or adapted to fit the task at hand. While data-driven methods offer the potential for more transparent and adaptable solutions, a fundamental challenge is the scarcity of accented data, which must be first addressed, then overcome. For the remainder of Section 2, I give a critical account of previous approaches belonging to both of these categories, emphasising their strengths and limitations in order to ultimately justify the methodological choices employed in my experiment.

2.2.2 Architecture-driven approaches

Hinsvark et al. (2021) review various technical approaches at improving the recognition of accented speech, identifying accent-specific modelling as an earlier key strategy. Accent-specific modelling involves a modular system comprising of an accent identification model, which is used to identify the

accent, and several accent-specific acoustic models which are used during inference depending on the label generated by the accent identification module. Although this strategy does perform well, not only does it need enough accented data to train each accent-specific model, but learning parameters for each individual model separately is costly – ideally, a single model should learn to perform well on both native and non-native speech.

Model generalisation techniques, such as multi-task learning (MTL) or domain expansion, offer promising solutions to this challenge. Multi-task learning allows models to share parameters across different tasks, so that the top layers remain accent-specific, while layers further down share acoustic parameters. Shor et al. (2019), for instance, address the task of improving the recognition accuracy for both dysarthric and foreign-accented speech by only fine-tuning the parameters of a particular subset of layers in the encoder of the RNN-T model they employ, though it remains unclear whether the accent features are truly separated from speaker identity. Overall, while this strategy does yield improved performance, it still relies on either an accent identifier or a manually-input accent ID, which makes it a rather indirect solution. Li et al. (2021) comparatively explore multi-task learning (MTL) and domain-adversarial learning (DAL) for accented English, using accent embeddings trained either in a supervised or in an unsupervised fashion. Their results show that DAL in combination with labelled embeddings promotes the learning of accent-invariant features, while MTL with unsupervised wav2vec2 embeddings perform best on unseen accents. Thus the need for accent labels can be overcome via unsupervised accent embedding learning, although overall the improvement seen even with the best experimental setting are rather incremental.

MTL has also been explored in comparison with transfer learning. Ghorbani and Hansen (2018), for instance, propose treating foreign-accented speech as an interpolation of L1 and L2, by training a model on native speech (e.g. Spanish, Hindi, English) and evaluating whether this improves the model's performance on accented combinations of those languages. Their experiments demonstrate that, while pre-training with native languages in this manner does improve ASR performance, MTL yields better results. Specifically, using native language data as a secondary task increases the model's tolerance to accented speech, with the primary task focusing on native English and the secondary task on native Spanish or Hindi.

Domain expansions strategies, on the other hand, improve a baseline model's performance on a new accent through regularisation or novel architectures, while maintaining the same performance on the previous set of accents – unlike fine-tuning, where the performance on the initial task is likely to degrade. For instance, Na and Park (2021) use a domain-adversarial neural network (DANN) for domain adaptation, aiming to minimise distributional differences between accented and non-accented speech. There are three subcomponents to the model; a feature extractor (CNN-based) trained on mel-spectrograms extracts relevant accent features from speech, a domain classifier (DNN-based) classifies the speech as either accented or not, based on the previously extracted features, and lastly a label predictor (CTC-based) predicts the final character labels. When tested on various English accents (Canadian, Australian, British, Indian), this modular approach improved the recognition performance across all accents, with the most significant improvement obtained when the most target accent data was used.

2.2.3 Data-driven approaches

Several recent studies have explored and evaluated data augmentation techniques aimed at improving the performance of ASR systems on non-native accents from a data perspective. Since one of the main

obstacles when it comes to accented speech is the scarcity of such speech samples, data augmentation is one of the main avenues of research when mitigating accent bias.

Zhang et al. (2023), for instance, evaluated several such methods in an attempt to address ASR bias against groups of non-standard speakers, including children, elderly people, non-native speakers, and people with speech impairments. The authors identify three main categories of data augmentation techniques based on how the original speech signal is modified. Speed and volume perturbation are the more common and 'traditional' data augmentation techniques in ASR, applied directly to the time domain; they involve re-sampling the original audio to alter its tempo and adjusting the volume. Feature warping and masking includes techniques such as SpecAugment (Park et al., 2019), where warping is applied directly to the feature inputs and masking is applied to blocks of frequency channels and blocks of time steps. Lastly, a more recent data augmentation method involves the use of perceptually-driven perturbations (e.g. pitch shifting, which changes the pitch contours while maintain the articulation pattern unchanged) or that of voice conversion models, which aim to create new articulation patterns altogether and thus yield significant added variation in the training data.

Fukuda et al. (2018) also examined several audio signal-level augmentation methods, such as noise addition, speed modification, and voice conversion; they find significant improvements in transcribing Latin American and Asian accents (data from several languages binned together), with speed modification surprisingly being the most effective, voice conversion providing some benefit, and noise addition actually degrading performance in comparison to the baseline. Importantly, their approach assumes the accent identity is known in advance and explicitly input, which is, however, not a realistic expectation for online implementation. More recently, experiments carried out by Zhang et al. (2023) find that combining multiple data augmentation techniques, such as SpecAugment with pitch and speed perturbation, yielded the best results. Additionally, using a small amount on non-native natural data to generate more artificial non-native data for training by means of voice converison also increased ASR robustness to foreign-accent-related pronunciation alternatives, reducing the WER.

Similarly, Zhang, Zhang, Halpern, et al. (2022) highlight advantages of voice conversion (VC) over classic time-frequency perturbation techniques such as SpecAugment and speed perturbation. This is likely due to the fact that, while the latter modify the speech signal directly and help with robustness in noisy environments, they do not address the intrinsic quality of accented speech. Voice conversion, on the other hand, preserves speech characteristics of the original data and generate new speech signals which are qualitatively different along a particular, controlled scale. The VC model utilised in the study by Zhang, Zhang, Halpern, et al. (2022) is AGAIN-VC (Chen et al., 2020), which is an autoencoder-based, cross-lingual voice conversion model that can be used to synthetically generate accented speech via transfer learning and domain-adversarial training. Experimental results show that using this model in a data augmentation approach to increase ASR robustness to non-native pronunciation led to the lowest WER on both investigated architectures, compared to more traditional data augmentation techniques. Moreover, the authors also note that bias mitigation is model-dependent, with the Transformer-based model showing improved performance compared to an RNN-based model, regardless of the data augmentation or training strategies used.

Especially in more recent studies, VC as a data augmentation tool has seen increased popularity in addressing various flavours of underresourced ASR settings, including foreign-accented speech. Some early attempts at VC go as far back as Zhao, Sonsaat, Levis, Chukharev-Hudilainen, and Gutierrez-Osuna (2018), where the authors aim to isolate and eliminate the features corresponding to a perceived foreign accent, thus converting non-native speech to sound like native speech. The authors achieve this by matching the frames of two speakers into a phonetic posteriogram based on

phonetically-informed similarity, obtaining a well-rated acoustic quality (measured with mean opinion score – MOS) and higher accent ratings compared to mapping based on acoustic similarity. More recently, Ding, Zhao, and Gutierrez-Osuna (2022) Accentron model is designed to generate accent-converted speech for any arbitrary L2 speaker, even unseen during training, and addresses two significant challenges in early foreign accent conversion: the need for speech data from each L2 speaker for training and the need for a separate models for each L1-L2 pair. To overcome this, the model utilises a speaker-independent acoustic model trained on L1 speech to extract bottleneck features and represent the linguistic content of the L1 utterances. Furthermore, its zero-shot nature also allows the model to synthesise speech for arbitrary non-native speakers, necessitating a strong accent-speaker features disentanglement.

Several recent studies have presented promising one-shot – or even zero-shot – accent conversion models (ACM) which are able to convert input native speech to foreign-accented speech while preserving speaker identity and linguistic content. Jia et al. (2023) propose such a zero-shot, reference free ACM, which preserves speaker identity features by using a timbre encoder from mel-spectrograms to model a speaker's timbre, producing natural-sounding speech (MOS 3.45), with perceptually salient accentedness (MOS 3.82) and reasonable speaker similarity (MOS 3.13). Jin et al. (2023) build upon this work by developing a more flexible zero-shot ACM that can convert an unseen speaker's utterances to multiple accents while preserving the original voice identity, using adversarial learning to disentangle accent-dependent features. Experiments on eight accents show high scores in audio quality (MOS 3.62), speaker similarity (MOS 4.05), and accent conversion, with synthesized samples often preferred for sounding "more accented" compared to the originals. Lastly, Melechovsky, Mehrish, Sisman, and Herremans (2022) introduce a novel framework for accented text-to-speech (TTS) using a conditional variational autoencoder. This TTS system synthesizes a selected speaker's speech in any desired accent without the need for reference audio once trained. Subjective MOS-based evaluations indicate that the perceived naturalness of the synthesized speech is similar to the original audio. The study also finds promising results for speaker identity retention and accent similarity, though it notes a trade-off between accent strength and identity preservation, highlighting the complexity of balancing these aspects since accent forms a part of a person's identity.

Overall, it seems that traditional data augmentation techniques such as speed and volume perturbation, as well as feature warping and masking, have shown some success in increasing ASR performance on the speech of underrepresented speaker groups. However, they often fall short in addressing the intrinsic qualities of accented speech. Studies on voice conversion models collectively highlight the potential, as well as the main challenges of (especially zero-shot) accent conversion; encoder architectures or domain-adversarial training have been successfully addressing the main challenge of teasing apart accent features from other aspects, such as speaker identity. At the same time, there is a persistent trade-off between perceived accentedness and speaker identity preservation in the final result. Moreover, it is often the case that promising zero-shot VC models do not have openly available code, which makes them impossible to directly use as a data augmentation tool. Nevertheless, voice conversion has proven to be an important data augmentation tool in mitigating accented speech bias. For example, Klumpp et al. (2023) use an accent conversion model to synthesize foreign-accented English speech, then use the generated data to train an ASR system on pronunciation alternatives; they find that the use of synthetic data significantly improves the model's performance on foreign-accented samples, though it's not clear whether the performance on native speech was affected by training the ASR model on accented speech from scratch.

In light of the limitations associated with accent conversion models mentioned above, in the

present study I investigate whether a state-of-the-art speech synthesis model can act similarly to a voice conversion approach in generating synthetic training samples. Specifically, I aim to find whether a speech synthesis model trained on native Dutch speech and fine-tuned on a small subset of single-speaker accented Dutch data can learn to reliably emulate accented Dutch speech, as well as whether such synthetic data can contribute to improving the performance of an ASR model on natural accented speech. If successful, this could overcome the issues of open-access availability or need parallel utterances associated with some VC models, as well as the data scarcity challenge in the area of accented speech recognition, because once a synthesis model learns to generate the desired speech profile, it can be employed for the synthesis of virtually unlimited amounts of data.

In sum, Section 2 highlights the complexities and persistent challenges associated with ASR systems in relation to accented speech. Despite significant advancements, speech recognition models continue to fail at closing the performance gaps across speaker subgroups. Although traditional approaches, including the use of balanced datasets or time-domain data augmentation techniques, offer some mitigation, they often fall short in addressing the nuanced acoustic shifts induced by foreign accents. Advanced techniques, such as multi-task learning, domain-adversarial training, or voice conversion models, show promise, but come with their own set of limitations, related to data scarcity and trade-offs between accent strength and speaker identity preservation.

The review further underscores the necessity for more innovative solutions that can effectively integrate and augment accented speech data, without being overly reliant on large, diverse datasets of accented speech, as these are often unavailable. In this context, the exploration of speech synthesis models as potential tools for generating synthetic accented speech presents a promising direction. This approach might help in closing the gap left by traditional data augmentation techniques without further complicating the pipeline, by increasing the amount of speech feature variations the model sees during training.

3 Methodology

3.1 Data

Three types of Dutch speech data are used for this experiment, as indicated in Table 1:

	Total duration	**No. of utterances**	**Used for**	**Source**
ND	14h 6m 40s	6494	Baseline and pre-training FS2	CSS10 Dutch
NA-D-test	27m 51s	125	Evaluating the performance of fine-tuned whisper models	HuggingFace
NA-D-train	1h 53m 25s	500	Fine-tuning FS2	
SA-D	12h 48m	5494	Fine-tuning whisper	HuggingFace

Table 1: Overview of data subsets

Specifically, the abbreviations refer to:

- **Native Dutch (ND) speech**: By 'native Dutch', I mean Dutch spoken by a native speaker. I use the CSS10 dataset of Dutch speech, comprising of 14 hours of read speech (single-speaker, male voice), specifically Jules Verne's novel *20.000 mijlen onder zee*.
- **Natural accented Dutch (NA-D) speech**: By 'naturally accented Dutch' I mean Dutch spoken by an L2 speaker of Dutch which is directly recorded (i.e. not obtained through speech synthesis or speech conversion). The NA-D dataset used in this study is a single-speaker, small dataset of accented Dutch speech. A detailed account of the relevant characteristics in the speaker profile can be found in Appendix A.
- **Synthetic accented Dutch (SA-D) speech**: By 'synthetic accented Dutch', I mean foreign-accented Dutch speech that is obtained synthetically, i.e. using a speech synthesis model fine-tuned to produce accented Dutch speech.

3.2 Models and training strategy

3.2.1 Whisper

OpenAI's Whisper model (Radford et al., 2022) performs speech recognition by making use of large-scale weak supervision during learning, and works across several languages and tasks (transcription and translation). Trained on over 680.000 hours of multilingual and multitask data, the authors claim the model does not need fine-tuning for down-stream tasks and thus has zero-shot transfer capabilities. Whisper uses an encoder-decoder Transformer architecture (Vaswani et al., 2023), processing input audio into mel-frequency cepstrums and passing them to the encoder; the decoder then predicts text captions conditioned on both audio features and previously predicted text, which means that the model is able to capture long-term dependencies and contextual information.

However, given the training strategy and the data sources used, it is not entirely transparent what data Whisper has seen during training, such as how many languages, or how many hours of each language. As the paper by Radford et al. (2022) mentions, whisper learns from a large and diverse set of audio-text data pairs from the internet, with a 'minimalist approach to data pre-processing' (p.2), which comprises of various speakers, languages, recording setups, and environments. Some amount of filtering was performed in an automated fashion, to remove poor quality transcriptions, such as cases where the audio and the transcript were not classified as containing the same language by a language detection system. This minimalist approach to data preprocessing is meant to allow Whisper to predict raw test, with no extensive standardisation, while enabling the construction of a large dataset without extensive time or human annotation resources.

3.2.2 FastSpeech2

FastSpeech2 (Ren et al., 2022) is a relatively recent speech synthesis model designed to speed up the synthesis process while maintaining a high quality in the final audio result. Compared to the previous version (FastSpeech – (Ren et al., 2019)), it eliminates the student-teacher training paradigm by learning directly from the ground truth data, thus improving duration prediction and information loss in mel-spectrogram generation. The FastSpeech2 model architecture includes an encoder, a variance adaptor, and a mel-spectrogram decoder, employing a feed-forward Transformer block for efficient processing. This leads to a significantly improved performance compared to both its predecessor and other models, while achieving a much faster inference time at the same time. Indeed, this was confirmed in my experiment, as generating over ten hours of speech and per-utterance spectrograms took around 40 minutes, which makes it significantly faster than real-time. MOS evaluations further underscore its state-of-the-art status, as FastSpeech2 achieves an overall MOS score of 3.83, compared to 3.6 for its older counterpart.

FastSpeech2 is crucial in this experimental study for obtaining the SA-D data subset. To this end, the speech synthesis model is first trained on the ND dataset for 600.000 steps, then fine-tuned on the NA-D-train dataset for an additional 100.000 steps. This yields a high perceptual quality of the synthesised audios, with a good level of accentedness and the speaker identity of the NA-D dataset. The synthesised samples can be found here. Further detail about the training process of the FastSpeech2 model can be found in Appendix C.

3.2.3 Trasnfer learning: Pre-training and fine-tuning

Transfer learning is a learning method in the field of machine learning which aims to train models to resolve a new task by leveraging an assumed similarity between datasets, model architectures, or the nature of the tasks from one (old) problem to another (new) one (Wang & Chen, 2023). This is useful for numerous scenarios, such as when properly annotated data is absent or when trying to avoid heavy-duty computations associated with training models from scratch. Pre-training and fine-tuning belong to the transfer learning paradigm and are aimed at adapting a set of parameters acquired on a previous related task, given a novel target dataset of limited size.

This learning paradigm is particularly suitable for the present experiment. Native speech has a fundamental similarity to foreign-accented data, due to the common underlying linguistic structure, with the latter displaying some acoustic shifts in the feature space compared to the former. It is also reasonable to expect that this underlying similarity might alleviate the 'forgetting issue' associated with transfer learning as well, which refers to a model losing its performance on the task learned dur-

ing pre-training at an accelerated rate due to its parameters quickly adapting to fit the newly presented data (Liu et al., 2024). The present experimental design takes this into account, as I additionally test whether a potential increased performance on accented speech after fine-tuning brings about a decrease in performance on native Dutch speech.

At the same time, overfitting is a common risk of using fine-tuning as a learning strategy. Overfitting occurs when the model learns the training data too closely, and thus loses its ability to generalise to new, unseen data. This is commonly reflected during training by diverging training loss values compared to validation loss values; more specifically, the training loss decreases (which means the model gets increasingly good at predicting the training data), while the validation loss increases (which means the model loses its ability to correctly predict unseen samples). An additional risk factor for overfitting in the present approach comes from the fact that the pre-trained Whisper checkpoints were trained on a large amount of domain-general data (160.000 hours of many languages, across two tasks), but fine-tuning happens on a comparatively very small set ($\tilde{1}0$ hours) of domain-specific (single-speaker accented Dutch). To prevent overfitting in my fine-tuning experiment, I implement parameter-efficient fine-tuning (PEFT) via low-rank adaption (LoRA), as introduced by Hu et al. (2021).

PEFT is a method first used for adapting pre-trained language models to specific tasks while minimising the computational and storage costs associated with retraining all of its parameters, given the fact that such models have millions, if not billions of parameters. LoRA (Hu et al., 2021) is one such approach, whereby pre-trained model weights are frozen and low-rank decomposition matrices are injected into each layer of the Transformer architecture. This significantly reduces the number of trainable parameters required for fine-tuning, leading to a faster training time and reduced GPU memory usage. Additionally, by focusing on low-rank updates of the weight matrices, LoRA prevents overfitting by ensuring that only the relevant and essential changes are made to the model parameters during fine-tuning, thus largely maintaining the generalisation capabilities of the pre-trained model.

3.3 Experimental conditions

3.3.1 Baseline

A significant body of previous literature reviewed above informs us that out-of-the-box ASR models perform very well on native speech data, but significantly worse on non-native (foreign-accented) speech data. Therefore, I compare the results of my approach to a baseline WER score of "out-of-the-box" whisper models (base, small, medium, and large) on both a subset of the ND data and a subset of NA-D data. A motivating factor behind this choice is to check whether the model that is widely considered to be the state-of-the-art in ASR at this point in time (Radford et al., 2022) still displays the strong bias against accented speech that has been previously documented in the literature.

3.3.2 Fine-tuning Whisper on synthetic data

The main research question investigated in this study is related to the feasibility of using synthetically-accented data to improve the performance of an ASR model on pronunciation variation in naturally-accented data (i.e. 'in the wild'). To this end, this experimental condition explores the performance of whisper models on naturally accented data (which reflects the foreign-accented speech 'in the wild') when fine-tuned on synthetically-accented data obtained from a FastSpeech2 model. Detailed fine-tuning specifications used can be found in Appendix C.

By exploring the quality of synthesized speech at several checkpoints along the fine-tuning process, it was observed that the best acoustic quality, speaker characteristics preservation, and reliable accent emulation occur after 100K steps of fine-tuning. Therefore, the SA-D dataset is obtained by performing inference on a subset of utterances that the model had not seen during training or fine-tuning, from the 500K checkpoint (i.e. FastSpeech2 trained for 400K steps on native speech, followed by 100K steps on custom, NA-D speech). Several relevant checkpoints, including the one used in the study, can be found here. An important advantage is the fact that, once FastSpeech2 learns how to produce accented speech, there is virtually no limit as to how much synthetic speech data can be produced using it, and so the data scarcity challenge can be overcome – quantitatively speaking. Nevertheless, it remains an open question whether a subset of synthetically accented data, no matter how large, can actually improve the ASR model's performance on natural speech samples of non-native Dutch.

4 Results

4.1 Baseline

To fairly measure the impact of my approach against the current state of ASR for non-native Dutch speech, I have set as a baseline the mean word error rates (WER) obtained when using Whisper (base, small, medium, large) on both accented data (NA-D) and native Dutch (ND) data. Figure 1 shows the mean word error rates (WER) on native Dutch speech compared to foreign-accented Dutch speech across whisper model sizes, indicating that each model performs significantly worse (specifically, has a higher mean WER) on foreign-accented data compared to native data – in line with Hypothesis 2.

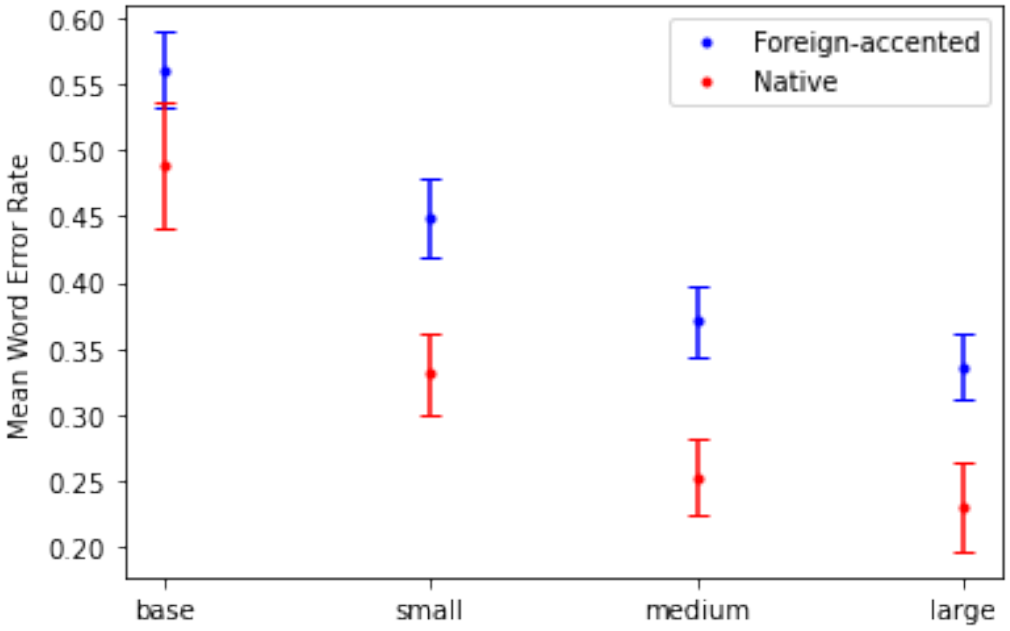


Figure 1: Mean WER: The performance of Whisper models **before fine-tuning** on native and natural accented Dutch speech.

		Mean WER	SE	*p*-value
base	**native**	48.9%	2.42%	$p = 0.013$ *
	accented	56%	1.45%	$p = 0.013$ *
small	**native**	33.1%	1.52%	$p = 1.616 \times 10^{-7}$ **
	accented	44.8%	1.52%	$p = 1.616 \times 10^{-7}$ **
medium	**native**	25.3%	1.48%	$p = 2.696 \times 10^{-8}$ **
	accented	36.9%	1.36%	$p = 2.696 \times 10^{-8}$ **
large	**native**	23%	1.71%	$p = 1.706 \times 10^{-6}$ **
	accented	33.5%	1.26%	$p = 1.706 \times 10^{-6}$ **

Table 2: Mean WER, standard error (SE) and *p*-values for Hypothesis 2 (baseline).

Furthermore, a pairwise Welch's t-test across model sizes confirms this empirical observation; Welch's t-test is suitable for comparing the means of two independent groups in the absence of an equal variance assumption and applies to this case because the utterances corresponding to each baseline are different, a case for which the Student's t-test lacks robustness. The null hypothesis (specifically, that WER means for each model size are equal for the native and accented speech data) can thus be rejected for all models (base: $p = 0.013$; small: $p = 1.616 \times 10^{-7}$; medium: $p = 2.696 \times 10^{-8}$; large: $p = 1.706 \times 10^{-6}$). In other words, a significant gap in performance on foreign-accented Dutch samples compared to the native Dutch counterpart persisted in my data across Whisper model sizes.

4.2 Fine-tuning Whisper

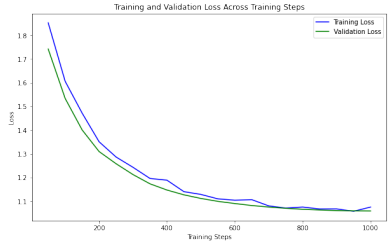
In the fine-tuned experimental condition, Whisper models (base, small, medium, and large) were fine-tuned on the SA-D dataset from their pre-trained checkpoints, using the training specifications found in Appendix B. The training and validation losses, as well as the WER on the validation set across training steps can be seen in Figure 2.

The **whisper-base** model underwent fine-tuning over 1000 training steps, proving to be the most stable model size during this process. Initially, the WER on the validation set was notably high at 71.44%, but it consistently decreased to 52.44% by the final step. Throughout the training, both the training loss and validation loss remained aligned, without divergence. The final checkpoint saved was the last one (1000).

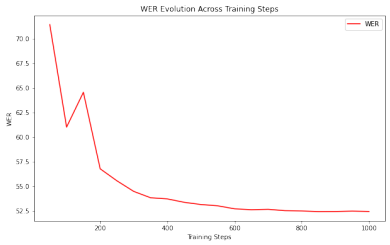
The **whisper-small** model underwent fine-tuning over 1000 training steps as well, but with different hyperparameter settings, although this model size displayed pronounced instability that was hard to mitigate despite numerous experiments. Initially, the WER on the validation set was very high at 77.9% and it decreases rapidly all the way to 29.92% by the final step, although the validation loss plateaus around the 200th training point despite the training loss continuing to go down. As this indicates that the model starts to overfit the training data around this point in time, the final saved checkpoint was the one at the 200th step, despite the fact that this is not associated with the lowest validation WER score. This is because a model that did not overfit the new data is expected to perform better on the NA-D test set, despite the higher validation WER.

The **whisper-medium**, fine-tuned over the same number of steps, started at a WER of 67.2% on the validation set, which decreases relatively smoothly all the way to 36.2% throughout training. At the same time, the training and validation loss values do not diverge drastically, thus indicating that no overfitting occurred when fine-tuning this model. The training loss, however, remains relatively high even at the end of the 1000th step, which indicates the model might be able to learn more from the data given a higher number of training steps or different training hyperparameters. Nevertheless, the 1000th checkpoint was saved at the end for this model.

Lastly, the **whisper-large** model was trained for 600 steps, as its increased number of parameters gives it a high capacity to learn from data, which means that it is particularly prone to overfit small datasets. Despite this measure, the large whisper model shows the most unstable behaviour during fine-tuning, as reflected by the validation WER curve, which does not stabilise across iterations. Moreover, due to time and computation constraints, a single hyperparameter configuration was tested for fine-tuning whisper-large; in other words, there are likely better fine-tuning settings which can be discovered through more extensive experimentation under more generous time conditions. Nevertheless, in the current setting, the best checkpoint to save seemed to be the 250th one, where the WER is the lowest and the training and validation loss curves are on a downward trend.

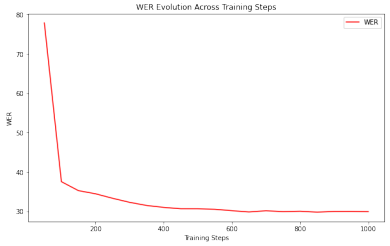


(a) Training and validation loss (whisper-base)



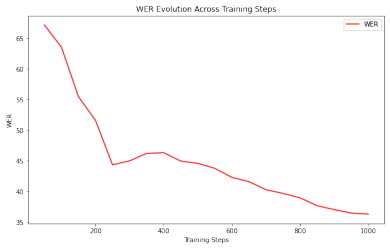
(b) Validation WER (whisper-base)

(c) Training and validation loss (whisper-small)



(d) Validation WER (whisper-small)

(e) Training and validation loss (whisper-medium)



(f) Validation WER (whisper-medium)

(g) Training and validation loss (whisper-large)

(h) Validation WER (whisper-large)

Figure 2: Evolution of whisper model fine-tuning across steps for different model sizes.

4.3 Fine-tuning results

The WER reported in the previous subsection were related to the unseen validation set of synthetic data, as part of the training process; the research questions in this study, however, is related to the performance of the fine-tuned Whisper models on natural accented Dutch speech (NA-D), which I have not discussed yet. Figure 3 and Table 3 show the mean WER scores obtained when using the fine-tuned whisper checkpoints to transcribe both native Dutch and natural accented Dutch.

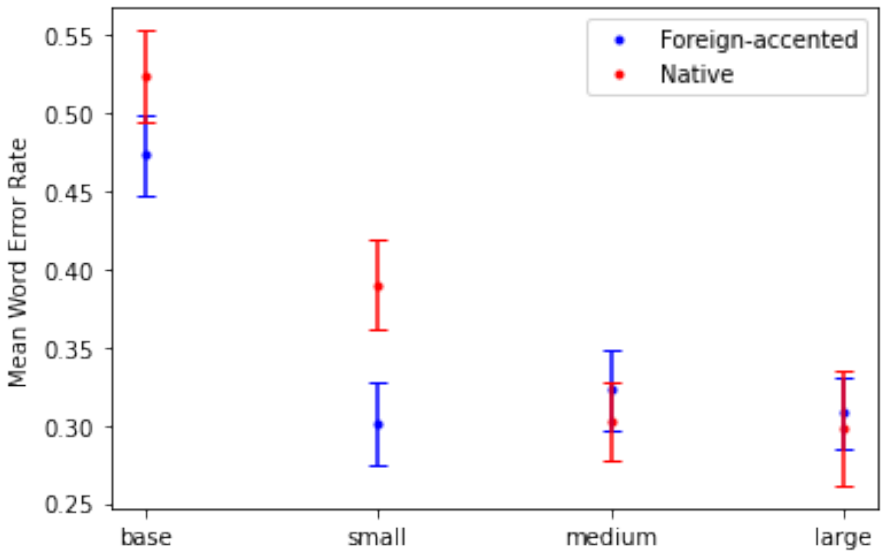


Figure 3: Mean WER: The performance of **fine-tuned** Whisper on native vs. natural accented Dutch speech.

		Mean WER	**SE**	***p*-value**
base	**native**	52.3%	1.49%	$p = 0.012$ *
	accented	47.3%	1.31%	$p = 0.012$ *
small	**native**	39%	1.46%	$p = 1.076 \times 10^{-5}$ **
	accented	30%	1.35%	$p = 1.076 \times 10^{-5}$ **
medium	**native**	30.3%	1.27%	$p = 0.277$
	accented	32.3%	1.32%	$p = 0.277$
large	**native**	29.8%	1.88%	$p = 0.653$
	accented	30.8%	1.17%	$p = 0.653$

Table 3: Mean WER, standard error (SE), and p-values for the performance of **fine-tuned Whisper** on native vs. accented speech.

The WER achieved by the fine-tuned models is significantly lower across accentedness conditions only for the whisper-base model ($p = 0.012$) and for the whisper-small model ($p = 1.076 \times 10^{-5}$). At the same time, by looking at the mean WERs displayed in Figure 3 above, it can be observed that gap is flipped; the fine-tuned base and small models performed significantly worse (i.e. had a

significantly higher WER) on *native* Dutch, and not on accented Dutch speech like in the baseline condition reported above (4). This finding is in line with the training observations reported above, especially in the case of the small model, where clear overfitting was observed. For the other two models, the mean WER difference between accented Dutch and native Dutch was reduced to statistical insignificance (medium: $p = 0.277$; large: $p = 0.653$).

However, this test alone does not directly answer all the initial research questions of the study. Before I proceed with discussing the final statistical results, I reiterate the initial hypotheses presented under Section 1.2, with the mention that the test for the second hypothesis is presented above under 1:

> **H1**: Fine-tuning the Whisper model on synthetic accented Dutch speech improves its performance on natural accented Dutch speech. In other words, the WER of the fine-tuned model on accented speech is significantly lower than the WER of the pre-trained, out-of-the-box Whisper on the same accented dataset.
>
> **H2:** Whisper still displays the persistent performance gap across accented and native speech, as documented in previous literature. In statistical terms, a significantly higher WER can be observed on the natural accented data compared to the WER on the native data before fine-tuning
>
> **H3:** The fine-tuned model's performance slightly goes down on native Dutch speech, compared to its pre-trained counterpart.

To test **Hypothesis 1**, I conducted a Welch's t-test between the performance of the fine-tuned model and that of the initial pre-trained model on natural accented speech, shown in Figure 4. This was meant to elucidate whether the fine-tuning process with synthetic speech truly helped the model perform better on natural accented Dutch.

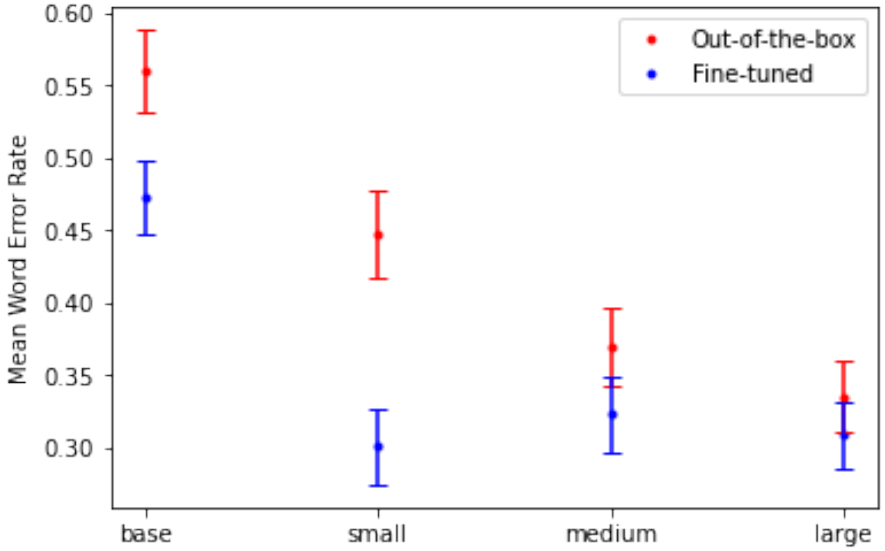


Figure 4: Mean WER: The performance of Whisper models **on natural accented speech** before vs. after fine-tuning.

		Mean WER	SE	*p*-value
base	**pre-trained**	56%	1.45%	$p = 1.506 \times 10^{-5}$ **
	fine-tuned	47.3%	1.31%	$p = 1.506 \times 10^{-5}$ **
small	**pre-trained**	44.8%	1.52%	$p = 7.521 \times 10^{-12}$ **
	fine-tuned	30%	1.34%	$p = 7.521 \times 10^{-12}$ **
medium	**pre-trained**	36.9%	1.36%	$p = 0.015$ *
	fine-tuned	32.3%	1.32%	$p = 0.015$ *
large	**pre-trained**	33.5%	1.26%	$p = 0.123$
	fine-tuned	13.1%	1.17%	$p = 0.123$

Table 4: Mean WER, standard error (SE), and p-values for the performance **on natural accented speech** before vs. after fine-tuning.

As observed in Figure 4, there is a systematic decrease in WER on accented speech for the fine-tuned model compared to its pre-trained counterpart, in line with the different fine-tuning behaviours as well. Both whisper-base and whisper-small perform significantly better on the task of transcribing natural accented speech after fine-tuning (base: $p = 1.506 \times 10^{-5}$; small: $p = 7.521 \times 10^{-12}$); interestingly, the large gap observed for the small model reflects a large improvement on natural accented speech, despite the fact that the fine-tuning behaviour indicated overfitting. Although its performance on the synthetic validation set hit a ceiling relatively fast, it seems the model still learned enough from this data to cause an important improvement on natural samples. The whisper-medium model also performs significantly better on accented speech after fine-tuning ($p = 0.015$), although the large model does not ($p = 0.123$), likely because of the underfitting it displays during fine-tuning, showing that the model did not learn enough during the process to make a significant difference.

The same statistical test was conducted to test **Hypothesis 3**, this time between the pre-trained and the fine-tune model's performance on native Dutch, in order to test whether the WER on native speech significantly increases in parallel with the improvement on accented speech.

		Mean WER	SE	*p*-value
base	**pre-trained**	48.9%	2.42%	$p = 0.231$
	fine-tuned	52.3%	1.49%	$p = 0.231$
small	**pre-trained**	33.1%	1.52%	$p = 0.005$ **
	fine-tuned	39%	1.46%	$p = 0.005$ **
medium	**pre-trained**	25.3%	1.48%	$p = 0.012$ *
	fine-tuned	30.3%	1.27%	$p = 0.012$ *
large	**pre-trained**	23%	1.71%	$p = 0.008$ **
	fine-tuned	29.8%	1.88%	$p = 0.008$ **

Table 5: Mean WER, standard error (SE), and p-values for the performance **on native speech** before vs. after fine-tuning.

As Figure 5 shows, this is true for all models except whisper-base (base: $p = 0.231$; small:

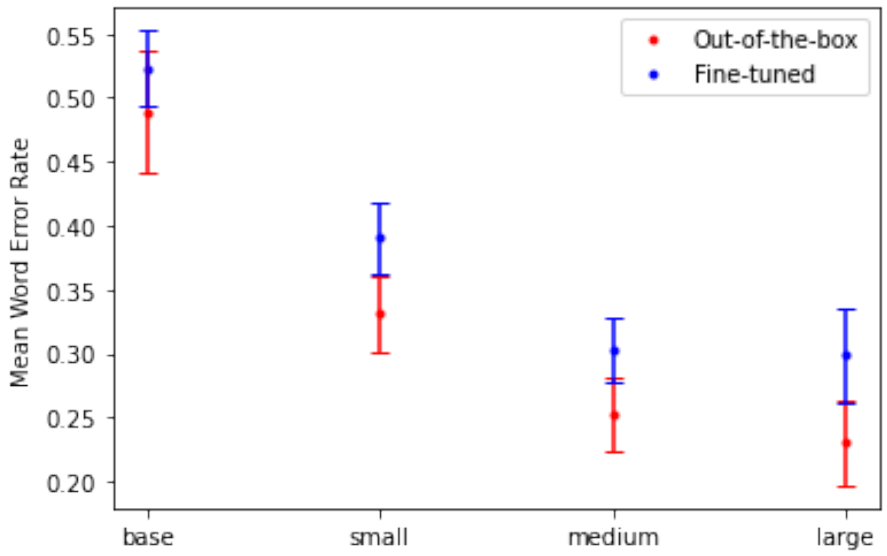


Figure 5: Mean WER: pre-trained ("out-of-the-box") vs. fine-tuned Whisper's performance on native Dutch.

$p = 0.005$; medium: 0.012; large: $p = 0.008$). In other words, for whisper-small, whisper-medium, and whisper-large, the WER for native Dutch speech was significantly higher compared to accented Dutch speech, indicating a decreasing performance after fine-tuning.

To sum up, the results of the experiments presented here with respect to the initial hypotheses are as follows:

1. **Hypothesis 1** is partially confirmed: The performance of smaller Whisper models increased significantly on accented speech, but this was not the case for the larger model; this is likely associated with the observed underfitting issue, which means the model did not learn enough from the synthetic data in order for its performance on natural accented Dutch to significantly increase.

2. **Hypothesis 2** is confirmed, providing additional motivation for the current study and contributing to the baseline of the experiments. Specifically, the reported tests show that even the latest most robust ASR system performed significantly worse on the accented Dutch speech samples compared to the native Dutch counterparts.

3. **Hypothesis 3** is partially confirmed: All model sizes except for whisper-base showed a systematic decrease in performance on native Dutch after fine-tuning, indicating that the additional learning process caused the models to slightly 'forget' how to perform the initial task.

The following section (5) contains a detailed discussion of these results, how they fit in the bigger picture of the field's literature, as well as what future directions of research they might open up.

5 Discussion

5.1 Baseline results

The baseline results indicate two trends, both in line with previous literature. Firstly, increased model size correlates with a decrease in WER for both native Dutch (ND) and foreign-accented Dutch (NAD) speech; Figure 1 highlights that larger whisper models consistently achieve lower mean error rates compared to their smaller counterparts. However, across all model sizes, there is a significant performance difference on non-native compared to native speech, with each model displaying notably higher mean WER on foreign-accented data.

While gaps in performance across foreign-accents have been previously documented in the literature (as reviewed here as well under 2), the results obtained here are somewhat surprising nevertheless, especially in terms of magnitude. For example, the Whisper Github page shows much lower WER scores for Dutch data from the Common Voice dataset (5.3% for whisper-large) and from the FLEURS dataset (5.2% for whisper-large). The present study has found comparatively higher WER for this model size, on both native (mean WER of 23% for whisper-large) and non-native (mean WER of 36% for whisper-large) Dutch speech from the datasets employed here.

One factor this could be attributed to is the register in the read speech data from the CSS10 corpus of Dutch speech. This dataset comprises of Jules Verne's book *20.000 mijlen onder zee*, which was written in 1870 and thus contains samples of older Dutch words and phrases. For example, upon some random manual inspection of the data samples, I have encountered relatively frequent uses of the word *gij*, which roughly corresponds to the old English pronoun *thou*, while at the same time sounding very similar to the common (modern) Dutch pronoun *hij* "he". A similar example of words prone to misrecognition is the name of one of the main characters in the novel, *Ned Land*, which sounds very similar to the Dutch endonym "Nederland" and was transcribed as such on most occasions, especially prior to fine-tuning. With more time at hand, such cases could perhaps be handled separately in a semi-automated fashion, so that the final transcripts are briefly post-processed before WER scores are computed. Additionally, further methods could be used to understand the exact cases that the model finds difficult to handle. For example, a phoneme-level confusion matrix could help better understand whether there are any particular phonemes in the accented dataset that the whisper model struggles to recognise, while a qualitative, manual inspection of a few random transcriptions could yield interesting insights that metric-driven approaches might not point out directly.

Nevertheless, the baseline findings underscore the persistent challenge in the field of ASR when it comes to accented speech, as even recent state-of-the-art models struggle to reliably handle accented speech. The factors discussed above probably affected the performance on both native and foreign-accented speech rather equally, so both associated WER increased, but should not have affected directly the gap between them which remains quite large. This reinforces the necessity of the present study, while also highlighting the need for continued research into accent bias and for dedicated attention and resources to develop more inclusive and accurate ASR systems.

5.2 Fine-tuning Whisper

The fine-tuning results across Whisper models reveal important insights into their performance, stability, and overall training behaviour, as well as their suitability for such a transfer learning paradigm. The **base** model, for example, shows good stability throughout the fine-tuning process, with all metrics (training loss, validation loss, validation WER) dropping consistently, yet the training and the

validation loss remain at high levels even at the end of the process, which is further reflected by the WER. Thus it seems like there is only so much that whisper-base can learn from the SA-D dataset. **Whisper-small** displays some common signs of overfitting, as its training loss continuously decreases while both the validation loss and the validation WER hit a ceiling relatively early, despite numerous re-runs and various hyperparameter adjustments. Conversely, **whisper-medium** is rather on the underfitting side, as reflected by the relatively high training loss at the end, as well as all other metrics that remain quite high. Lastly, the **whisper-large** model shows overall increased instability in the validation WER, although the training and validation loss curves show that might be learning some limited patterns in the data, up to a certain point. Better fine-tuning results might be achieved with a higher number of iterations in combination with learning rate scheduling and a larger batch size, although this requires highly performing computational resources; though it might also be the case that the largest version of the whisper model is simply not a good choice for a fine-tuning paradigm.

Overall, these practical insights highlight that fine-tuning whisper on a custom speech dataset might be an art of its own. A delicate balance could be observed between model capacity, training stability, and the ability to generalise to an unseen subset of synthetic data. It is not entirely clear, however, how much of the observed instability or limited learning capacity stems from the nature of the synthetic data, and how much is due to hyperparameter settings used in training. Before making further use of Whisper in a similar transfer-learning paradigm, a more detailed study into its fine-tuning behaviour is worth conducting. For example, a systematic way of testing the model's sensitivity (i.e. how much its performance is affected by different training settings given the same input data) and its stability (i.e. how much its performance changes when the training settings remain unchanged, given data subsets with different underlying distributions) across sizes could attest whether fine-tuning is a feasible learning strategy for Whisper at all.

5.3 General discussion

The performance of fine-tuned Whisper models on NA-D speech compared to ND speech reveals the main findings of my thesis and underscores the previously documented challenges of accented speech recognition. Firstly, the present experiments further confirmed that a significant gap in performance across accented vs. native Dutch speech still persists even for the latest and most robust ASR architectures, a gap which cannot be attributted entirely to the data quality issues discussed above. Thus the present paper reinforces the need for continued efforts in closing this performance gap across dialectal and foreign-accent speech variation, which continues to be an issue to date.

The main research question of the present study was whether synthetically generated samples of accented Dutch speech could help Whisper models become more robust to accent-characteristic pronunciation alternatives in natural speech, similarly to how previous studies have shown that voice conversion can be a promising data augmentation tool in this regard. The smaller models showed significant performance improvements in the fine-tuned condition; the medium model improved as well, though not as drastically as the smaller counterparts, while the large model did not show a significant improvement, likely due to the underfitting issue. However, this increase in performance on accented speech also meant a significant performance decline on native speech for almost all models, indicating that even parameter-efficient fine-tuning techniques such as LoRA cannot fully overcome the forgetting issue associated with fine-tuning.

Overall, it could also be observed that seeing the available Whisper model sizes as increasingly bigger pearls along a necklace is admittedly a naive representation. As their vastly different and differently unstable fine-tuning behaviour demonstrates, it is reasonable to assume that each model's

increasing parameter number has more complex consequences in terms of their capacity, complexity, their ability to learn further, as well as what kinds of data they would need to do so. For example, both whisper-base and whisper-large, the two poles of the size spectrum, displayed some hints of underfitting the new dataset in the current hyperparameter setting; however, an reasonable intuitive observation would be that the base model might benefit from further training, whereas it is unclear that the large model would. The increased instability of the latter, paired with its immense capacity, might contribute to the ceiling effect observed.

Ultimately, the approach I present here, with identical fine-tuning settings, does not seem like a viable method of mitigating bias against accented Dutch speech. Rather than closing the performance gap across accentedness, the current results indicate that this particular fine-tuning approach slightly reduced and flipped the bias, disfavouring native speech. Apart from further studies into Whisper's transfer learning behaviour and capabilities, one could also wonder whether the synthetic data I used was of sufficient quality. To the human ear, the samples seem of sufficient quality, albeit it is quite recognisable as synthetic speech; however, to a speech recognition model, the vocoded metallic noise or pitch jumps might represent excessive noise that impedes learning. Looking into different speech synthesis models might be a viable alternative.

5.4 Limitations

One practical limitation of this study is the hardware used. To fine-tune large whisper models effectively, access to an A100 GPU is essential, due to its ability to accommodate a good batch size. However, during the time when the practical experiments were carried out, getting access to this powerful GPU via the Google Colab resource allocation system was most of the time impossible, so that all fine-tuning job were carried out on the L4 GPU, which offers 22.5GB of GPU RAM. While this was generally good enough for whisper-base, whisper-small, and whisper-medium, a significant reduction in batch size was required for whisper-large in order for the training process to take place at all. Moreover, unlike the other model sizes, for which the required resources allowed for several runs until a good training parameters combination was found, the whisper-large fine-tuning process was only run once; thus there might be training parameter configurations Although I attempted to mitigate the batch size of 4 by increasing the gradient accumulation steps to 4, the fine-tuning performance was likely sub-optimal due to these memory constraints.

Several limiting factors can be identified in terms of the datasets used as well. The CSS10 dataset, used for training FastSpeech2, as well as as a basis for the custom NA-D dataset, consists of read speech from a relatively old novel published in 1870. This dataset contains words and phrases which are no longer common in contemporary Dutch, which likely contributed to some extent to the mismatch between the WER reported by the authors on more recent speech data (such as the Common Voice corpus) and the significantly higher scores observed in this study, even on native Dutch speech. Since whisper was primarily trained on much more recent data, it likely struggled with older word version or did not follow the older orthography rules which exist in the original, ground truth text samples, which affected the main evaluation metric. This limitation should be explicitly addressed if this dataset is to be used in future similar experiments.

Additionally, the NA-D subset used in this experiment is a single-speaker dataset, arguably reflecting the real-world trade-off between low-resource availability and speaker diversity. It is unclear whether it would be more beneficial to have a larger amount of single-speaker data or rather a collection of multi-speaker data with fewer speech samples per participant. Unfortunately, the experiment I introduce in this thesis is not able to contribute significantly to this discussion, as the accented speech

subset is a single-speaker one. It is technically possible to use FastSpeech2 to synthesise speech in as many vocal identity flavours as the number of speakers seen in training; however, this method for generating synthetic speech data does not rely on speaker identity - accent disentanglement, which means that it is not intuitively transparent what type and strength of perceived accent a multi-accent, multi-speaker implementation would yield.

Lastly, striking a good balance between underfitting and overfitting during fine-tuning Whisper models was a persistent challenge. As the figures of losses and WER rates across steps in (figure) indicate, some model sizes overfit earlier than others. This underscores the challenges associated with fine-tuning such large models like Whisper, which are prone to overfit on smaller, domain-specific datasets due to their extensive learning capacity. A dataset such as the synthetic accented Dutch subset used in this experiment seems to not provide enough learning diversity to promote good generalisation, with the large model's increased variance causing rapid overfitting. Moreover, the complex interaction of numerous parameters create a less stable training environment; for example, some experimentation with hyper-parameter tuning showed an increased sensitivity to learning rate changes, with adjustments of one order of magnitude greatly impacting learning, despite the implementation of warm-up steps. Overall, simpler ASR architectures might be a better choice for fine-tuning strategies, as the present experiments indicate the Whisper's high capacity might not work well with the reduced sample complexity in the SA-D dataset.

6 Conclusions. Future directions

This study looked into the feasibility of using synthetic accented data from a speech synthesis model as training material for the Whisper automatic speech recognition model in order to reduce its bias against foreign-accented Dutch speech.

First, the baseline performance on ”out-of-the-box” Whisper on both accented and native Dutch confirmed that a significant performance gap exists, even for one of the most recent and robust speech recognition models to date. This gap is evident from the increased WER on foreign-accented Dutch, which aligns with previous literature in the field and reinforces the persistent issue of accent bias in state-of-the-art ASR.

Fine-tuning the Whisper models on synthetically generated accented Dutch yielded mixed results. Smaller models appeared more stable in achieving an improved performance after fine-tuning compared to the larger counterparts, though they were still highly susceptible to overfitting. Moreover, the ”forgetting” phenomenon associated with transfer learning paradigms was not entirely avoided, despite efforts to implement optimized fine-tuning via low-rank adaption.

Overall, the results I have presented make it clear that mitigating bias against foreign-accented speech is not a trivial challenge in ASR and, while synthetically generated accented samples can be a promising data augmentation tool to overcome the data scarcity issue associated with this challenge, the acoustic quality and perceived accentedness of the final samples should be ensured before starting the fine-tuning process. Despite previous success obtained by using synthetic speech samples from accent conversion models as additional accented data to improve cross-accent robustness in ASR systems, the present experiment did not yield results that align with this success. This reinforces the need to better understand Whisper’s capacity for and behaviour during transfer learning, as well as to further explore whether the synthetic samples obtained from FastSpeech2 are comparable to those from voice conversion models in a subjective listening test.

Nevertheless, voice conversion remains a strong tool in data-driven approaches to bias mitigation for foreign-accented speech, as evidenced by numerous previous experiments in this direction, while the suitability of a fine-tuning paradigm in reducing Whisper’s performance gap across accentedness does not seem as viable based on the current results. While using similar approaches to generate significantly more (e.g. thousands of hours) synthetic data might work, as they could achieve an amount comparable to the pre-training sample of the investigated model, a more promising future direction might be to focus on the *quality* of the synthetic samples and ensure the least amount of vocoding errors and noise. Similarly, a systematic review, as well as a practical evaluative implementation, of the existing open-source voice conversion modles might help in better understanding the current state of research in this direction, as well as what the options and their corresponding limitations are. Lastly, it seems that different training strategies apart from fine-tuning might yield better results, especially in the case of very large and highly complex models such as Whisper, as their large capacity to fit a variety of functions does not work well with the limited data resources of accented speech.

References

Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167639307000404 doi: https://doi.org/10.1016/j.specom.2007.02.006

Chen, Y.-H., Wu, D.-Y., Wu, T.-H., & Lee, H.-y. (2020, October). *AGAIN-VC: A One-shot Voice Conversion using Activation Guidance and Adaptive Instance Normalization.* arXiv. Retrieved 2024-05-22, from http://arxiv.org/abs/2011.00316 (arXiv:2011.00316 [cs, eess])

Cucchiarini, C., Driesen, J., Hamme, H. V., & Sanders, E. (2008). Recording speech of children, non-natives and elderly people for hlt applications: the jasmin-cgn corpus.

Ding, S., Zhao, G., & Gutierrez-Osuna, R. (2022, March). Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*, *72*, 101302. Retrieved 2024-05-13, from https://linkinghub.elsevier.com/retrieve/pii/S0885230821001029 doi: 10.1016/j.csl.2021.101302

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024, March). Towards inclusive automatic speech recognition. *Computer Speech & Language*, *84*, 101567. Retrieved 2024-05-13, from https://linkinghub.elsevier.com/retrieve/pii/S0885230823000864 doi: 10.1016/j.csl.2023.101567

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021, April). Quantifying Bias in Automatic Speech Recognition. Retrieved 2024-02-01, from http://arxiv.org/abs/2103.15122 (arXiv:2103.15122 [cs, eess])

Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023, October). Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (pp. 146–151). Bucharest, Romania: IEEE. Retrieved 2024-05-16, from https://ieeexplore.ieee.org/document/10314895/ doi: 10.1109/SpeD59241.2023.10314895

Fukuda, T., Fernandez, R., Rosenberg, A., Thomas, S., Ramabhadran, B., Sorin, A., & Kurata, G. (2018, September). Data Augmentation Improves Recognition of Foreign Accented Speech. In *Interspeech 2018* (pp. 2409–2413). ISCA. Retrieved 2024-05-13, from https://www.isca-archive.org/interspeech_2018/fukuda18_interspeech.html doi: 10.21437/Interspeech.2018-1211

Ghorbani, S., & Hansen, J. H. (2018, September). Leveraging Native Language Information for Improved Accented Speech Recognition. In *Interspeech 2018* (pp. 2449–2453). ISCA. Retrieved 2024-05-13, from https://www.isca-archive.org/interspeech_2018/ghorbani18_interspeech.html doi: 10.21437/Interspeech.2018-1378

Hinsvark, A., Delworth, N., Del Rio, M., McNamara, Q., Dong, J., Westerman, R., ... Jette, M. (2021, June). *Accented Speech Recognition: A Survey.* arXiv. Retrieved 2024-02-08, from http://arxiv.org/abs/2104.10747 (arXiv:2104.10747 [cs, eess])

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *Lora: Low-rank adaptation of large language models.*

Jia, D., Tian, Q., Peng, K., Li, J., Chen, Y., Ma, M., ... Wang, Y. (2023, August). *Zero-Shot Accent Conversion using Pseudo Siamese Disentanglement Network.* arXiv. Retrieved 2024-04-16, from http://arxiv.org/abs/2212.05751 (arXiv:2212.05751 [eess])

Jin, M., Serai, P., Wu, J., Tjandra, A., Manohar, V., & He, Q. (2023, October). *Voice-preserving Zero-shot Multiple Accent Conversion.* arXiv. Retrieved 2024-02-02, from http://arxiv.org/

abs/2211.13282 (arXiv:2211.13282 [cs, eess])

Klumpp, P., Chitkara, P., Sarı, L., Serai, P., Wu, J., Veliche, I.-E., ... He, Q. (2023, March). Synthetic Cross-accent Data Augmentation for Automatic Speech Recognition. Retrieved 2024-02-01, from http://arxiv.org/abs/2303.00802 (arXiv:2303.00802 [cs, eess])

Kudina, O. (2024, February). *Voice-based Interfaces: Diversity and inclusion considerati* [Keynote]. Beeld en Geluid, Hilversum, The Netherlands.

Li, J., Manohar, V., Chitkara, P., Tjandra, A., Picheny, M., Zhang, F., ... Saraf, Y. (2021, October). Accent-Robust Automatic Speech Recognition Using Supervised and Unsupervised Wav2vec Embeddings. Retrieved 2024-02-01, from http://arxiv.org/abs/2110.03520 (arXiv:2110.03520 [eess])

Liu, S., Keung, J., Yang, Z., Liu, F., Zhou, Q., & Liao, Y. (2024, February). *Delving into Parameter-Efficient Fine-Tuning in Code Change Learning: An Empirical Study.* arXiv. Retrieved 2024-06-04, from http://arxiv.org/abs/2402.06247 (arXiv:2402.06247 [cs])

Martin, J. L., & Wright, K. E. (2023, August). Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, *44*(4), 613–630. Retrieved 2024-05-16, from https://academic.oup.com/applij/article/44/4/613/6901317 doi: 10.1093/applin/amac066

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. interspeech 2017* (pp. 498–502). doi: 10.21437/Interspeech.2017-1386

Melechovsky, J., Mehrish, A., Sisman, B., & Herremans, D. (2022, November). *Accented Text-to-Speech Synthesis with a Conditional Variational Autoencoder.* arXiv. Retrieved 2024-02-02, from http://arxiv.org/abs/2211.03316 (arXiv:2211.03316 [cs, eess])

Na, H.-J., & Park, J.-S. (2021, September). Accented Speech Recognition Based on End-to-End Domain Adversarial Training of Neural Networks. *Applied Sciences*, *11*(18), 8412. Retrieved 2024-02-08, from https://www.mdpi.com/2076-3417/11/18/8412 doi: 10.3390/app11188412

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019, September). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019* (pp. 2613–2617). Retrieved 2024-05-22, from http://arxiv.org/abs/1904.08779 (arXiv:1904.08779 [cs, eess, stat]) doi: 10.21437/Interspeech.2019-2680

Quamer, W., Das, A., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2022, September). Zero-Shot Foreign Accent Conversion without a Native Reference. In *Interspeech 2022* (pp. 4920–4924). ISCA. Retrieved 2024-05-09, from https://www.isca-archive.org/interspeech_2022/quamer22_interspeech.html doi: 10.21437/Interspeech.2022-10664

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December). *Robust Speech Recognition via Large-Scale Weak Supervision.* arXiv. Retrieved 2024-05-21, from http://arxiv.org/abs/2212.04356 (arXiv:2212.04356 [cs, eess])

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). *Fastspeech 2: Fast and high-quality end-to-end text to speech.*

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019, November). *FastSpeech: Fast, Robust and Controllable Text to Speech.* arXiv. Retrieved 2024-06-05, from http://arxiv.org/abs/1905.09263 (arXiv:1905.09263 [cs, eess])

Schuurman, I., Schouppe, M., Hoekstra, H., & Van der Wouden, T. (2003). Cgn, an annotated corpus of spoken dutch. In *Proceedings of 4th international workshop on linguistically interpreted*

corpora (linc-03) at eacl 2003.

Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., ... Matias, Y. (2019, September). Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Interspeech 2019* (pp. 784–788). ISCA. Retrieved 2024-05-13, from https://www.isca-archive.org/interspeech_2019/shor19_interspeech.html doi: 10.21437/Interspeech.2019-1427

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). *Attention is all you need.*

Wang, J., & Chen, Y. (2023). *Introduction to Transfer Learning*. Singapore: Springer.

Zhang, Y., Herygers, A., Patel, T., Yue, Z., & Scharenborg, O. (2023, December). *Exploring data augmentation in bias mitigation against non-native-accented speech.* arXiv. Retrieved 2024-04-16, from http://arxiv.org/abs/2312.15499 (arXiv:2312.15499 [eess])

Zhang, Y., Zhang, Y., Halpern, B., Patel, T., & Scharenborg, O. (2022, September). Mitigating bias against non-native accents. In *Interspeech 2022* (pp. 3168–3172). ISCA. Retrieved 2024-05-03, from https://www.isca-archive.org/interspeech_2022/zhang22n_interspeech.html doi: 10.21437/Interspeech.2022-836

Zhang, Y., Zhang, Y., Patel, T., & Scharenborg, O. (2022, September). Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems. In *1st Workshop on Speech for Social Good (S4SG)* (pp. 15–19). ISCA. Retrieved 2024-05-03, from https://www.isca-archive.org/s4sg_2022/zhang22_s4sg.html doi: 10.21437/S4SG.2022-4

Zhao, G., Sonsaat, S., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2018, April). Accent Conversion Using Phonetic Posteriorgrams. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5314–5318). Calgary, AB: IEEE. Retrieved 2024-02-02, from https://ieeexplore.ieee.org/document/8462258/ doi: 10.1109/ICASSP.2018.8462258

Appendices

A NA-D subset: Speaker profile

Personal information:

- Age at the moment of recording: 25 years and 9 months.
- Gender: female.
- Place of birth: Romania.
- Current residence: The Netherlands.
- Education level: higher education (graduate).

Linguistic background:

- Native language: Romanian (Daco-Romanian dialect).
- Non-native languages (in order of age of acquisition):
  - German:
    - age of acquisition: 7 years of age.
    - current proficiency level: CEFR B2.
    - context acquisition: formal education (primary and middle school with native German profile).
    - current daily use and exposure: minimal to none.
  - English:
    - age of acquisition: 10 years of age.
    - current proficiency level: CEFR C2.
    - context of acquisition: formal education, English media content and popular culture; over the previous 3 years: immersion, English-taught higher education programmes.
    - current daily use and exposure: every day, in formal and informal interactions, in higher education and academia.
  - Dutch:
    - age of acquisition: 23 years of age.
    - current proficiency level: CEFR A2/B1.
    - context of acquisition: formal education, immersion, self-study.
    - current daily use and exposure: daily, but limited.

B Whisper fine-tuning specifications

All the details regarding the training parameters used in fine-tuning each Whisper model can be found in the corresponding notebooks in the accompanying Thesis repository. Not all Whisper model sizes were fine-tuned with the same training configuration. In this Appendix, I give an account of the most important parameters and the reasoning behind the choice.

- **Learning rate**: Base, Small, and Medium used a learning rate of 10^{-5}; Large was trained at a slightly lower rate: 10^{-6}. Smaller models have fewer parameters and can converge faster with a lower risk of overfitting, while a smaller learning rate for the large model ensures a more stable training across its extensive network of parameters.
- **Batch size**: The batch size is the number of training samples fed to the model at once (in one iteration). A larger batch size usually helps keep the stability of gradient estimates, but is more computationally costly (memory-wise); a smaller batch size, however, can affect the gradient estimates and sloe down computation. Whisper also allows for setting the gradient accumulation steps, which means that gradients are accumulated over several mini batches before weights are updated. Decreasing batch size (and thus reducing memory requirements) can be paired with an increase in gradient accumulation points, which simulates a larger batch size. This was employed especially for fine-tuning the large model.
- **Number of epochs** represents the number of complete passes through the datasets. I have used the option of specifying the maximum number of steps, to get a finer-grained and more direct evaluation insight into the training process.

C FastSpeech2 training specifications

The CSS10 dataset of Dutch speech comprises of pairs of audio files (.wav format) and their correpsonfing transcriptions (.lab format). FastSPeech2 requires training data that is phoneme-aligned in order to extract duration information and train a duration predictor. In order to get from text-audio pairs to phoneme-aligned .TextGrids, the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) was used, which aligns text-audio pairs using a pronunciation dictionary of words in the target language and their corresponding phonetic transcription. Given time-aligned data, the speech can be prepared for training using the FastSpeech2 preprocessor, which extracts and normalises audio features.

For **pre-training**, the short-time Fourier Transform (STFT) parameters are set at a filter length of 1024, a hop length of 256, a window length of 1024, and the generated mel-spectrograms have 80 channels with a 0-8000 frequency range. The model specifications used include a transformer with 4 encoder layers (2 encoder heads, 256 encoder units), a convolution filter size of 1024, and kernel sizes of [9, 1]. The dropout rate is set to 0.2 for both the encoder and the decoder. Pitch and energy are quantized linearly with 256 bins. The mel-spectrogram is synthesized into the final audio via a HiFi-GAN vocoder with a maximum sequence length of 1000. The model is trained using these specifications for 600K steps, then it is **fine-tuned** on the NA-D dataset for an additional 100K steps using the same configurations.