university of groningen

campus fryslân

Exploring the Potential of Accent Conversion Techniques to Enhance Fairness in Language Assessment

Chenyu Li

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Exploring the Potential of Accent Conversion Techniques to Enhance Fairness in Language Assessment

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr.Matt Coler (Voice Technology, University of Groningen)
with the second reader being
xxxxx (Voice Technology, University of Groningen)

Chenyu Li (S-2615401)

June 11, 2024

Acknowledgements

I would like to extend my heartfelt gratitude to all those who have supported and guided me throughout the course of my master's program and the completion of this thesis.

I am deeply grateful to my supervisor, Dr.Matt Coler, for his unwavering support and guidance throughout the writing process. Your insightful feedback, encouragement, and expertise have been invaluable in shaping this work.

I also want to express my profound appreciation to my family, whose belief in my potential has been a constant source of motivation and strength. Your love and support have been my bedrock throughout this journey.

To my classmates and friends, thank you for your help and companionship. Your willingness to assist me without hesitation has made this experience all the more rewarding.

Lastly, I would like to express my gratitude to myself for the dedication and hard work over the past year. The countless late nights spent in front of the computer, exploring a brand new world, have been truly rewarding. I am amazed by looking back at the way I have gone so far..

Abstract

In the context of global mobility, ensuring equitable language proficiency assessments is crucial for fair immigration and integration policies. This thesis investigates the feasibility of using machine learning to neutralize Indian accents in English speech to enhance objectivity in language proficiency evaluations. The primary aim is to determine whether machine learning can effectively neutralize accents in English spoken by speakers whose native language is Hindi, thereby addressing identity anonymity in language test settings.

Existing foreign accent conversion (FAC) models are predominantly speaker-dependent, trained on datasets from specific speakers, and only effective for those individuals. The models that claim to work for unseen speakers typically involve a complicated pipeline structure, high data requirements, and are not easy to implement. This research aims to develop a speaker-independent model by training on a diverse dataset of Indian-accented English speakers. By doing so, it seeks to create a generalized accent conversion model with a simple structure that can be applied broadly, setting a precedent for extending this approach to other accents and thereby broadening the inclusivity of linguistic applications.

Applications of FAC include computer-aided language learning and entertainment, such as movie dubbing. However, the impact of FAC on language assessment has seldom been discussed. This study addresses a significant gap in language proficiency assessments, where the influence of accents on evaluation outcomes remains a challenge. Despite the recognition of accent-related issues by major testing organizations, explicit measures to mitigate their impact on scoring are lacking. Through innovative approaches to neutralize foreign accents in spoken language evaluations, this research aims to ensure fair and unbiased assessments for individuals from diverse linguistic backgrounds. By identifying and addressing these challenges, the study contributes to the advancement of equitable evaluation practices in multicultural societies.

Key words: Accent Neutralization, Accent Modification, Language Proficiency Assessment, Linguistic Inclusivity, Language Proficiency Assessment

Contents

1 Introduction

In the increasingly interconnected global landscape, the ability to communicate effectively in English has become a critical skill, influencing opportunities in education, employment, and immigration. However, traditional language proficiency assessments often fail to account for the biases introduced by non-native accents. These biases can lead to unfair evaluations, disproportionately disadvantaging individuals from diverse linguistic backgrounds.

Educational impacts: In educational settings, language proficiency assessments are pivotal for student placement, progression, and access to opportunities. For example, major language testing organizations like the International English Language Testing System (IELTS) strive to ensure their assessments are internationally oriented and free from accent bias (IELTS Asia, n.d.). Similarly, Cambridge English language assessment also notes that while a perfect English accent is not required, clarity in pronunciation is essential (Cambridge English, 2013). Nevertheless, the subjective nature of accent perception means that raters might still be unconsciously influenced by a candidate's accent, leading to inconsistent scoring. Despite these guidelines, explicit measures to mitigate the impact of accents on scoring may remain inadequate. Accents can inadvertently influence the perceived competence of non-native speakers, leading to biased scores that do not accurately reflect their true language abilities. By developing a speaker-independent machine learning-based accent conversion model, this research aims to neutralize such biases, ensuring that assessments are based on actual language proficiency rather than accent characteristics. This can lead to fairer placement decisions, more accurate assessments of student progress, and better educational outcomes for non-native speakers.

Immigration impacts: Language proficiency tests are often a key component of the immigration process, influencing decisions about residency and citizenship. For instance, in the Dutch citizenship exam, candidates' pronunciation is expected to be clear enough for understanding, even with a foreign accent (Kerkhoff, Poelmans, de Jong, & Lennig, 2005). Accents can skew these assessments, potentially leading to unfair outcomes that impact an individual's ability to immigrate or integrate into a new country. An accent-neutralizing model in these assessments can help ensure that applicants are evaluated on their language proficiency, promoting greater fairness and inclusivity in the immigration process.

Other societal impacts: Beyond education and immigration, the implications of this research extend to various sectors where effective communication is crucial. For example, some institutions, such as the German Federal Office for Migration and Refugees (BAMF), have started developing accent-detecting algorithms to verify someone's place of origin via their accent. These techniques have long been controversial and have been doubted due to the potential of violating human rights protections (AlgorithmWatch, 2021). In scenarios where people do not want to be identified and seek to increase privacy protection, accent conversion models could also be helpful. By converting accents, individuals can maintain their privacy and reduce the risk of being unfairly targeted or discriminated against based on their speech. Furthermore, this research contributes to the broader goal of promoting linguistic inclusivity and equity, fostering a more inclusive society where individuals are judged based on their abilities.

1.1 Research question

The persistence of these challenges highlights a critical gap in language assessment practices: the need for objective methodologies that can neutralize accent biases. Addressing this gap is essential to ensure fair and unbiased evaluations for individuals from diverse linguistic backgrounds. This recognition of the need for more equitable assessment practices forms the basis of the current research. Existing foreign accent conversion models are predominantly speaker-dependent, requiring extensive specific user speech data and parallel reference speech to achieve satisfactory results. The models that claim to work for unseen speakers typically involve a complicated pipeline structure, have high requirements for training data, and are not easy to implement. Building on this understanding, the core objective of this study is to explore the potential of a speaker-independent machine learning-based accent conversion model with a simple structure to enhance fairness in language proficiency assessments. The central research question driving this investigation can be formulated as follows:

> **Can a speaker-independent, simple-structured machine learning-based accent conversion model be developed to neutralize accents in English speech, thereby improving the fairness and objectivity of language proficiency assessments?**

To answer this overarching question, the study will focus on three subquestions:

- Is the converted speech still perceived as accented by listeners?
- Do the converted recordings receive significantly different scores in an experimental language test setting compared to the original recordings?
- Can the existing speaker-dependent accent conversion model be adjusted to accommodate non-specific speakers?

The hypotheses underlying this research are twofold:

- An accent conversion model can be effectively developed to neutralize target accents and conceal speaker identity in English speech, thereby enhancing the fairness of evaluations. This model is essential for application in language assessments to mitigate accent-related biases.
- A speaker-independent, simple-structure machine learning-based model can be developed to ensure fair language assessments.

For this study, the Indian accent was chosen due to the availability of relevant data. It is anticipated that the model will lead to measurable improvements in evaluation outcomes by minimizing accent-related biases. If the proposed method to generalize the conversion model fails, or if the converted speech does not result in significant changes in assessment scores, the hypotheses will be falsified. This outcome would suggest the need for further refinement of the model or the exploration of alternative approaches.

1.2 Structure of the thesis

The structure of the thesis is the following: subsection 1.1 introduces the research question posed along with a hypothesis on the outcome of the research. Section 2 provides an extensive literature review that frames the research question and hypothesis in the state-of-the-art. In section 3, the methodology is covered and the underlying models used are explained. Then, section 4 describes the experimental setup developed to answer the research questions and validate the hypothesis. Section 5 describes the results obtained and compares them to the baseline. In section 6, I discuss the previously-mentioned results in detail. Lastly, section 7 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

2 Literature Review

This section is dedicated to providing a comprehensive review of the existing research pertaining to accent conversion in speech processing, with a specific focus on developing a simple-structural speaker-independent accent conversion model to neutralize Indian accents in English speech. By conducting a thorough and critical analysis of the literature in this field, this review aims to offer valuable insights into its potential for enhancing fairness and objectivity in language proficiency assessments.

2.1 literature search methodology

The literature search for this study was conducted to ensure a comprehensive review of existing research on accent impact and foreign accent conversion. I employed a systematic approach to identify and select relevant literature, focusing on the latest advancements and influential works in the field. The following databases were utilized for the literature search:

- **I**EEE Xplore
- **W**eb of Science (WoS)
- **G**oogle Scholar

The search terms were grouped according to the topics they are related to. The topics and their corresponding keywords are as follows:

- **Accent impact:** rater-effect, accent bias, accent perception, linguistic inclusivity, language proficiency evaluation;
- **Foreign accent conversion techniques:** accent neutralization, accent modification, accent conversion, speaker-embedding, voice conversion;

To streamline the paper selection process and ensure relevance and quality, the following criteria were applied:

1. Topic Relevance: Papers were organized based on their relevance to specific topics and keywords. Papers not directly related to foreign accent conversion or accent impact were excluded to maintain coherence.
2. To ensure the inclusion of the most recent research, the selection criteria differed based on the category of the study. Only papers in linguistic studies published from 2000 onwards were considered. Only papers in Machine learning studies published from 2019 onwards were reviewed in detail. This criterion was applied to reflect the rapid advancements and methodologies in accent modification and voice technology within the field of machine learning.
3. To prioritize the most influential works, I selected the top 20 articles according to their relevancy/number of citations.

The literature search involved several steps to identify and select the final set of papers:

1. Initial Search: Conducted an initial search in the specified databases using the identified keywords. This search yielded a broad set of papers.

2. Screening for Relevance: Screened the titles and abstracts of the retrieved papers to assess their relevance to the research topics. Papers that did not align with the focus on accent impact and foreign accent conversion were excluded.

3. Application of Inclusion/Exclusion Criteria: Applied the inclusion criteria of publication date and influence (top 20 articles by relevancy/number of citations) to narrow down the selection.

4. Full-Text Review: Conducted a full-text review of the remaining papers to ensure their pertinence and quality. Papers that met all criteria were included in the final set for the literature review.

By following this systematic approach, the literature search aimed to ensure the inclusion of the most pertinent and up-to-date literature directly related to accent conversion. This methodology aligns with the research objectives and scope, providing a robust foundation for the study.

Based on the general topics, subsection 2.2 discusses the literature on the impact of accents. Moving towards current practices in accent conversion, subsection 2.3 introduces the major methods used.

For simplicity and readability, table 1 provides a full list of references appended with some notes, sorted by order of appearance in the following subsections of the literature review.

2.2 Research on accent impact

Rater effects (i.e., the construct-irrelevant variation in scoring due to raters' backgrounds) have long been a debatable topic in language assessment. Some researchers delving into the fields of speech processing and perception have discovered that familiarity with a specific accent can significantly boost an individual's ability to comprehend speech with that accent (Ockey & French, 2016). Bradlow and Bent (2008) observed that listeners perceptually adapt to foreign-accented speech upon repeated exposure, which enhances their ability to discriminate and identify sounds, suggesting a facilitating effect of accent familiarity on accent identification and listeners' ratings of speaker intelligibility, regardless of the speaker's baseline level of intelligibility. Building upon this study, Winke and Gass (2012) investigated whether raters' knowledge of test takers' first language (L1) affects how the raters orient themselves to the task of rating oral speech. The study involved native English-speaking raters with Spanish, Chinese, or Korean as their L2. They assessed non-native speech samples from Spanish, Chinese, or Korean L1 speakers. Results showed that raters who were familiar with the test taker's accent tended to give higher scores, while those who were not familiar tended to give lower scores. Moreover, the level of familiarity with an accent has been linked to the formation of stereotypes about foreign accents among listeners (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002) and their preferences for certain accents (Scales, Wennerstrom, Richard, & Wu, 2006).

However, some research on the impact of accent familiarity on listeners' judgments of speech has yielded inconsistent findings. For instance, Kennedy and Trofimovich (2008) reported no significant

disparity in the judgments of comprehensibility and accentedness of nonnative speakers between listeners who had been exposed to L2 speech and those who had not. Further research by Xi and Mollaun (2011) did not find a significant difference between the numerical ratings assigned by raters with and without familiarity with the speakers' accents as well. Similarly, B. H. Huang, Alegre, and Eisenberg (2016) found no significant difference in these judgments, yet a majority of the raters self-reported that their familiarity with an accent influenced their evaluations, potentially leading to a more lenient assessment of speakers with familiar accents. Recent research on Finnish accents has indicated that the impact of accents on oral examination scores is a case-by-case issue, with some accents receiving greater bias than others (Ahola & Halonen, 2021).

The reviewed literature highlights the significant role of accent familiarity in speech comprehension and evaluation, revealing both facilitating effects and potential biases. These findings are critical to my research question. The inconsistencies and biases identified in the literature underscore the need for an accent-neutralizing model that can mitigate these effects.

2.3 Research on accent conversion

Foreign accent conversion (FAC) is a specialized application of voice conversion (VC) focused on transforming accented speech from a non-native speaker into native-like speech while maintaining the speaker's identity.

FAC has seen significant evolution in its methodologies, shifting from complex articulatory trajectory analysis to more streamlined approaches. Initially, efforts focused on intricate details such as lip and tongue movements alongside vocal tract length normalization to transform non-native speech patterns into native-like ones(Aryal & Gutierrez-Osuna, 2015) (Felps, Geng, & Gutierrez-Osuna, 2012) (Aryal & Gutierrez-Osuna, 2016). These early methods, while insightful, were cumbersome and required extensive data on speaker's articulatory movements.

The advent of deep learning has ushered in a new era for FAC, with recent studies leveraging phonetic posteriorgrams (PPGs) and textual data. PPGs, which represent the posterior probability of speech frames belonging to predefined phonetic units, have emerged as a key feature due to their ability to encapsulate the linguistic and phonetic content of speech. Zhao, Ding, and Gutierrez-Osuna (2019) proposed a methodology(figure1) which involved using an acoustic model trained on a native speech corpus to extract speaker-independent phonetic posteriorgrams (PPGs). A speech synthesizer was then trained to map the PPGs from non-native speakers to their corresponding spectral features. Finally, a high-quality neural vocoder, WaveGlow, was used to convert the spectrogram into the raw speech signal. This approach has become a cornerstone in the field of accent conversion. However, this method has several limitations. Firstly, it functions more as a voice conversion model rather than an accent conversion model. During the conversion process, reference speech from a native speaker is required to extract PPGs as input for the model. The output is native-sounding speech with the voice identity of the non-native speaker. Additionally, for each non-native speaker conversion, both the synthesizer and the neural vocoder need to be retrained. Moreover, the acoustic model used for PPG extraction must be trained on a large corpus; in Zhao's experiment, it was trained on 960 hours of native speech data. Subsequent enhancements have built upon Zhao's baseline model to further refine accent conversion processes. For instance, Li et al. (2020) contributed

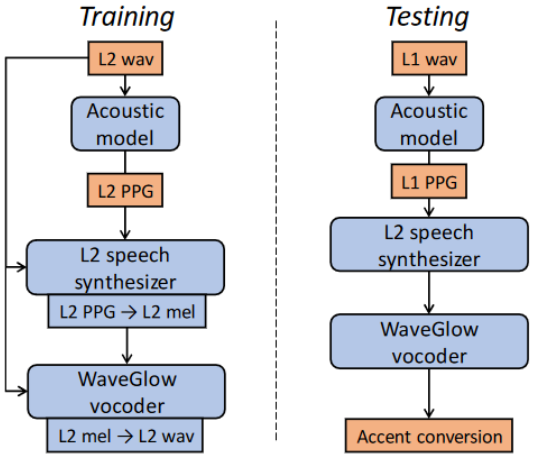


Figure 1: Overall workflow of FAC proposed by Zhao et al. (2019)

to improving the robustness of accent conversion models by integrating a reference encoder within an end-to-end text-to-speech system. This system uses multi-source information, including acoustic features from native speech and linguistic data, to elevate both the quality and authenticity of the converted speech. This method not only preserves the speaker's identity but also significantly enhances the native-likeness of the accent. Additionally, Zang, Xie, and Weng (2022) introduced the use of concentrated attention mechanisms within a modified Tacotron2 framework. This approach aims to improve the alignment between input phonetic sequences and mel-spectrograms, allowing for more precise accent modifications by focusing on the most relevant features within sequences. Consequently, the model's ability to mimic native speech nuances is enhanced. While these methods improved performance by adjusting the attention mechanism and incorporating additional references to modify the accent, they did not fundamentally alter the general three-component framework architecture. As a result, they still have limitations in practical applications.

Recent efforts in Foreign Accent Conversion (FAC) models have focused on developing reference-free methods. These approaches aim to achieve accent conversion without using native speech during the inference phase, thereby broadening the application scope of the model. W. Huang and Toda (2023) summarize the primary models and specific implementation methods for FAC and evaluate the performance of these models through experimental assessments. The study introduces and compares three main FAC models:

- **latent space conversion (LSC):** Quamer, Das, Levis, Chukharev-Hudilainen, and Gutierrez-Osuna (2022) introduce a novel zero-shot foreign accent conversion model(Figure 2) that can transform speech from previously unseen non-native (L2) speakers to sound as if it were produced by a native (L1) speaker.The model presented in the paper consists of 5 independent components, each responsible for a different aspect of the Foreign Accent Conversion (FAC) process:
  - Acoustic Model: This model takes an utterance from a speaker, whether a native (L1) or non-native (L2) speaker, and generates a Phonetic Posteriorgram (PPG). The PPG represents the posterior probability of each frame belonging to a set of predefined phonetic

units, capturing the linguistic content of the utterance in a speaker-independent manner. The model is implemented using Kaldi and trained on the Librispeech corpus.

- Speaker Encoder: This component is designed to capture the unique voice identity of a speaker. Trained as a speaker verification model, it produces a fixed-dimension embedding vector from a given utterance that represents the speaker's identity.
- Accent Encoder: Similar in architecture to the Speaker Encoder, the Accent Encoder is trained to recognize and capture the accent features of a speaker. It is used to obtain accent embeddings that characterize the specific accent of the speaker.
- Translator Module: This module comprises a sequence-to-sequence (seq2seq) model that utilizes Phonetic Posteriorgram (PPG) features from an L2 speaker's utterance along with accent embeddings from the Accent Encoder. It is trained to generate PPG features that would typically be produced by an L1 speaker, effectively translating the linguistic content from a non-native accent to a native one. However, in the actual experiment, the authors used bottleneck features (BNFs) instead of PPGs. BNFs are derived from the output of the last hidden layer of the acoustic model, which is the layer just before the final softmax layer. While BNFs contain similar linguistic information as PPGs, they have much lower dimensionality (256 vs. 6,024 for Senone-PPGs), making them more efficient for the model.
- Synthesizer Model: Also based on a seq2seq model, the Synthesizer takes the bottleneck features and speaker embeddings as inputs to synthesize a Mel-spectrogram for any given speaker. This Mel-spectrogram reflects the voice quality of the original speaker but with the accent characteristics of a native speaker.

Although the authors claim their model to be state-of-the-art, W. Huang and Toda (2023) evaluation shows that its performance is no better than that of the other two methods. However, it is the only model that achieves speaker-independence by utilizing PPGs and converting between non-native and native speakers in the latent space, as its name suggests. Despite its advantages, the implementation is highly complex due to its pipeline structure. The process involves first training the acoustic model on a large corpus of native speech, followed by training the speaker encoder and accent encoder. Additionally, training the synthesizer requires a parallel native and non-native speech dataset.

- **Synthetic Target Generation (STG):** In their effort to eliminate the need for native reference utterances in initial accent conversion models, Zhao, Ding, and Gutierrez-Osuna (2021) proposed a novel approach to foreign accent conversion (FAC) that functions without requiring reference native speech during inference. The proposed system leverages an acoustic model (AM) to generate speaker-independent (SI) speech embeddings for input utterances, whether they are from native (L1) or non-native (L2) speakers. The training process of the model consists of two main stages(Figure 3). In the initial step, the model uses a conventional FAC procedure to create a set of golden-speaker utterances (L1-GS). These golden-speaker utterances are produced by converting L2 speech to have native-like pronunciation while maintaining the voice identity of the L2 speaker. Essentially, they serve as the target utterances for the next step. Once the golden-speaker utterances are generated, the pronunciation-correction model is trained to map L2 utterances to these L1-GS targets. In short, this method employs a conventional model to synthesize target speech, which retains the L2 speaker's voice but adopts a

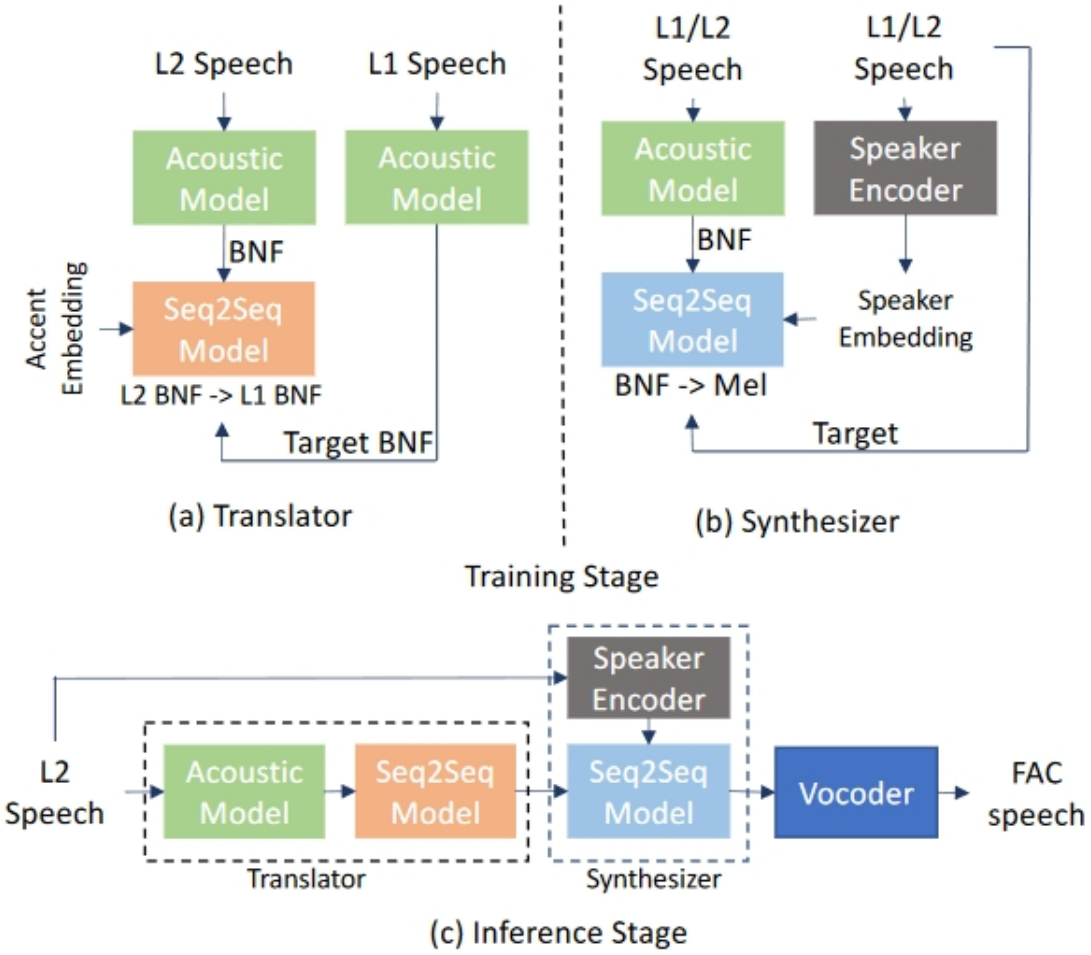


Figure 2: LSC model structure (Quamer et al., 2022)

native accent. This synthesized speech (L2 voice, L1 accent) is then used as training data for the model to learn how to convert accent while maintain the voice identity. This method still has several notable limitations. To train the acoustic model for speech embedding extraction, a large corpus is required. Since speaker embedding is part of the speech embedding dimension, the acoustic model must be trained on a dataset that covers a wide variety of speakers. The process requires the generation of golden speaker utterances for each L2 speech input. This implementation also involves a complex pipeline structure, requiring separate training phases for the acoustic model, speaker encoder, and accent encoder. Each component must be finely tuned and integrated, which can be technically challenging and resource-intensive.

- **Cascade Method:** Adapting the two-stage paradigm for preserving speaker identity in dysarthric voice conversion (DVC) proposed by W. Huang, Kobayashi, and Peng (2021), it has been found that this model is also effective for the task of accent conversion. The first stage uses a sequence-to-sequence (seq2seq) model to convert the input L2 speech into that of a reference L1 speaker. The authors employ a Transformer-based model named Voice Transformer

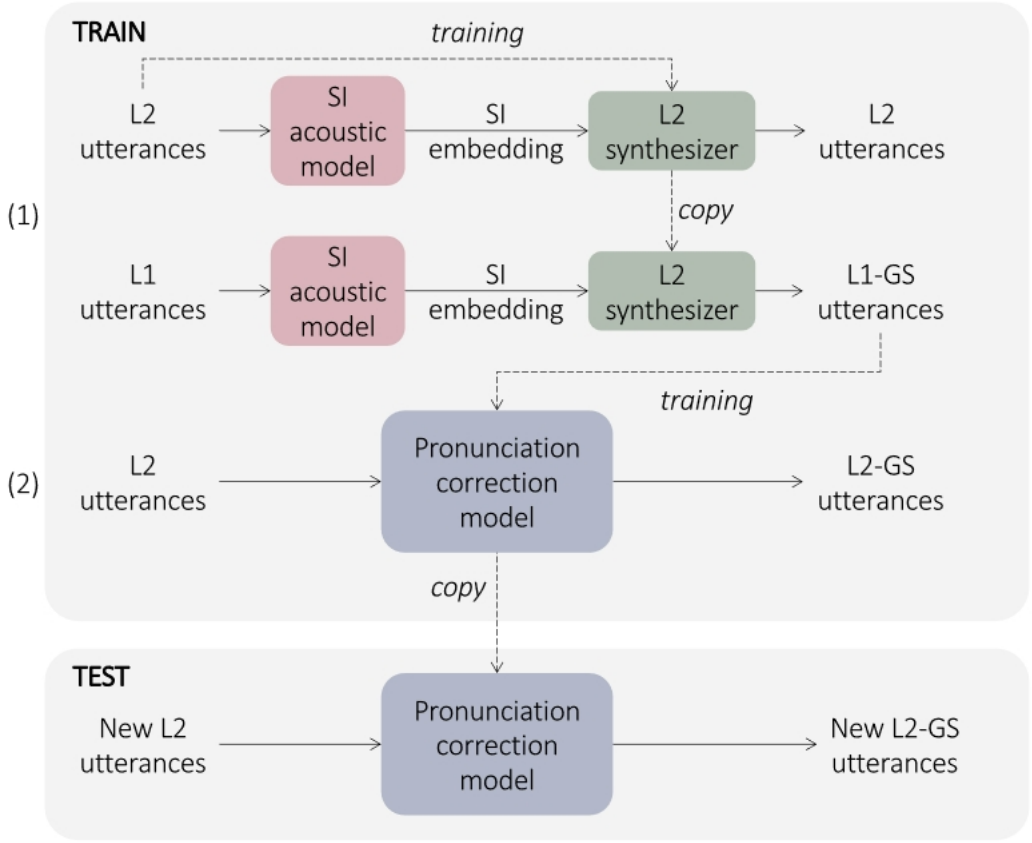


Figure 3: LSC model structure (Zhao et al., 2021)

Network (VTN) for this purpose, leveraging its ability to convert suprasegmental information and improve speech naturalness and intelligibility through parallel training. The second stage involves a frame-wise, nonparallel VC model realized with a variational autoencoder (VAE). This model takes the converted speech with the reference speaker's identity and restores the L2 speaker identity. The VAE is designed to change only time-invariant characteristics (such as speaker identity) while preserving time-variant characteristics (such as pronunciation), thus maintaining the improved speech quality. The method relies on a parallel dataset for training, where corresponding L2 and L1 speech pairs are needed. This requirement can be challenging as collecting parallel data from non-native and reference speakers is time-consuming and expensive. The performance of the model heavily depends on the choice of reference speakers. Variability in speaking rates, F0 patterns, and other speech characteristics among reference speakers can significantly affect the conversion quality and consistency. Selecting the optimal reference speaker requires careful consideration and potentially extensive trial and error.

Overall, the literature underscores a clear progression from labor-intensive articulatory analysis to data-driven, deep learning-enhanced methods that significantly simplify and improve the FAC process. The innovative use of PPGs and seq2seq models has set the stage for more natural and effective accent conversion techniques, moving the field closer to the goal of native-like speech synthesis for non-native speakers.

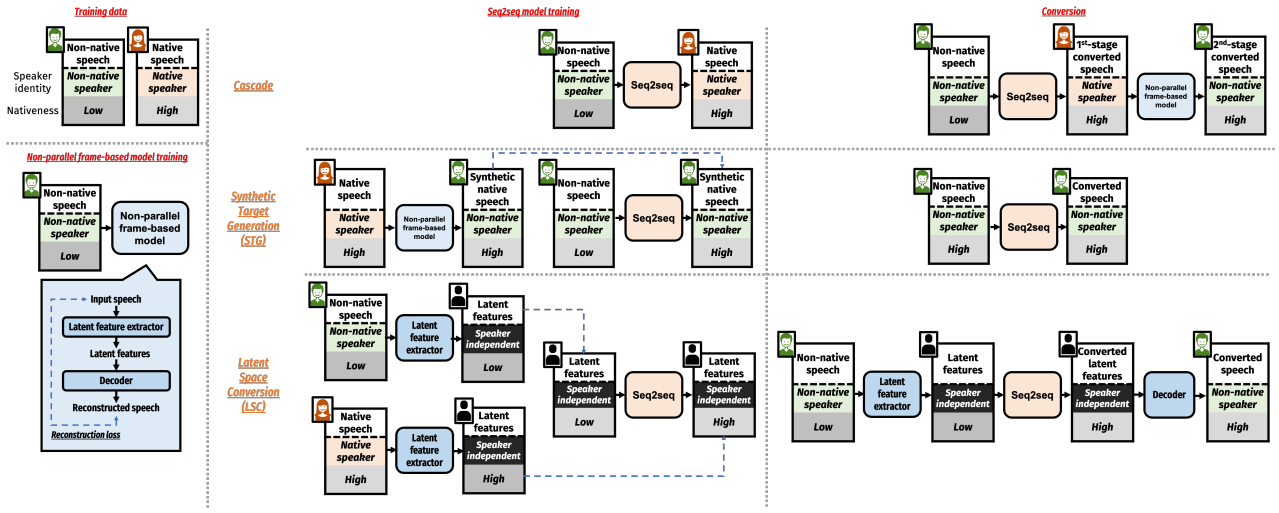


Figure 4: Comparison of FAC methods (W. Huang & Toda, 2023)

Figure 4 illustrates the three primary methods used for foreign accent conversion: latent space conversion (LSC), synthetic target generation (STG) and cascade. According to the conclusions from "Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion," the performance of these three models, when trained on the same dataset, did not show significant differences in effectiveness. Each model has its own strengths and weaknesses, and the choice of model may depend on specific application requirements and constraints (W. Huang & Toda, 2023).

In summary, the existing literature on accent impact and accent conversion provides valuable insights into the complexities and biases present in language proficiency assessments. Debate on rater effects underscores the need for objective assessment methodologies. Studies on accent conversion techniques, including recent advancements in reference-free approaches, lay the groundwork for innovative solutions to these biases. However, gaps remain in developing a speaker-independent model with a simple structure that can generalize across diverse speakers, maintain speech integrity, and be easily implemented.

Table 1: List of references for subsections 2.1-2.3, summarized

Reference	Brief description	Subsection
Ockey and French (2016)	From one to multiple accents on a test of L2 listening comprehension	2.2
Bradlow and Bent (2008)	Perceptual adaptation to non-native speech	2.2
Winke and Gass (2012)	The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation	2.2
Major et al. (2002)	The effects of nonnative accents on listening comprehension: Implications for ESL assessment	2.2
Scales et al. (2006)	Language learners' perceptions of accent	2.2
Kennedy and Trofimovich (2008)	Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context	2.2
Xi and Mollaun (2011)	Using raters from India to score a large-scale speaking test	2.2
B. H. Huang et al. (2016)	A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment	2.2
Ahola and Halonen (2021)	'Broken Finnish': Speaker L1 and its recognition affecting rating in National Certificates of Language Proficiency test in Finnish	2.2
Aryal and Gutierrez-Osuna (2015)	Reduction of nonnative accents through statistical parametric articulatory synthesis	2.3
Felps et al. (2012)	Foreign Accent Conversion Through Concatenative Synthesis in the Articulatory Domain	2.3
W. Huang and Toda (2023)	Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion	2.3
Zhao et al. (2019)	Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams	2.3
Zang et al. (2022)	Foreign Accent Conversion using Concentrated Attention	2.3
Li et al. (2020)	Improving Accent Conversion with Reference Encoder and End-To-End Text-To-Speech	2.3
W. Huang et al. (2021)	A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion	2.3
Zhao et al. (2021)	Converting Foreign Accent Speech Without a Reference	2.3
Quamer et al. (2022)	Zero-Shot Foreign Accent Conversion without a Native Reference	2.3
W. Huang and Toda (2023)	Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion	2.3

3 Methodology

In this chapter, I will outline the methodology employed to address the research question and validate the hypothesis regarding the neutralization of accents in speech for language assessments. First, in subsection 3.1, I will discuss the Foreign Accent Conversion (FAC) method that will be utilized in this study. Next, subsection 3.2 will provide an overview of the datasets used for training and testing the models, ensuring a diverse representation of accents. Following that, subsection 3.3 will delve into the adaptation of cascade models, elaborating on the specific model used in this study. Subsection 3.4 will then elaborate on the evaluation methods and metrics employed, with a particular focus on the word error rate (WER) as a key measure of accuracy. Finally, in subsection 3.5, I will reflect on the ethical considerations inherent in this research, including issues related to privacy, consent, and potential biases in the datasets and evaluation processes.

3.1 Selection of the Model-Cascade

This subsection discusses the rationale for choosing the cascade method for foreign accent conversion (FAC) in this study. The cascade method is a two-stage process that first maps accented speech to native-like speech and then restores the speaker's identity. The model consists of two primary components:

- Stage 1: A sequence-to-Sequence (seq2seq) model converts the input accented speech into a non-accented speech of a reference speaker. The seq2seq model used is based on the Transformer architecture, capturing both local and global dependencies in speech through multi-head self-attention layers. It takes log mel spectrograms as input and outputs converted counterparts with improved naturalness and intelligibility. Modern seq2seq models, often equipped with an attention mechanism to implicitly learn the alignment between the source and output sequences, can generate outputs of various lengths. This ability makes the seq2seq model a natural choice to convert duration in VC. In addition, the F0 contour can also be converted by considering F0 explicitly (e.g, forming the input feature sequence by concatenating the spectral and F0 sequences) (Tanaka, Kameoka, Kaneko, & Hojo, 2019) or implicitly (W. Huang, Hayashi, Wu, Kameoka, & Toda, 2020). Despite the decreased accentedness, this stage also results in a change of speaker identity to that of the reference speaker.

- Stage 2: A non-parallel frame-wise variational autoencoder (VAE) model restores the speaker identity of the accented speech while maintaining the accent style of the converted speech from Stage 1. The VAE model takes the converted speech from the seq2seq model as input. It modifies only the time-invariant characteristics (e.g., speaker identity) while preserving the time-variant characteristics (e.g., pronunciation).

Here, I explain why this method is particularly suitable for the goal of neutralizing accents in speech for language assessments.

The primary goal of the experiment is to neutralize accents in speech to enhance the fairness and objectivity of language proficiency assessments. By removing the accent, I aim to eliminate bias based on the speaker's accent, ensuring that evaluations are based solely on linguistic content. Additionally, maintaining identity anonymity can further enhance fairness, as it prevents raters from

being influenced by the speaker's voice characteristics.

The cascade method's structure is relatively simple, relying heavily on training with parallel data. The second stage, which restores the original speaker's voice identity could be omitted to maintain identity anonymity. This flexibility makes the cascade method ideal for applications where preserving or anonymizing identity is optional. The cascade method has been successfully applied in related tasks, such as dysarthric voice conversion, demonstrating robustness and effectiveness. W. Huang et al. (2021) showcased the method's ability to handle complex speech transformations while maintaining high levels of intelligibility and naturalness.

Furthermore, the model utilizes mel-spectrograms for training, which is solely based on the acoustics rather than a specific language, making it language-independent. Compared to the other 2 methods, it does not require the pre-training of a robust language-dependent acoustic model to extract language-dependent features, such as PPGs or BNFs. This simplifies the operational process and reduces the dependency on a specific linguistic environment, making the model more versatile and easier to implement.

3.2 Dataset

The datasets are L2-ARCTIC(Zhao, Sonsaat, Silpachai, et al., 2018), CMU_ARCTIC database and the NITK(National Institute of Technology Karnataka)-IISc Multilingual Multi-accent Speaker Profiling (NISP) (Kalluri, Vijayasenan, Ganapathy, Rajan, & Krishnan, 2020).

CMU_ARCTIC database is a US English single speaker databases designed for unit selection speech synthesis research. L2-ARCTIC is a speech corpus of non-native English intended for research in voice conversion, accent conversion, and mispronunciation detection. Each speaker recorded approximately one hour of read speech from CMU's ARCTIC prompts.

The NISP is an open-source non-native English speech corpus designed specifically for speaker profiling, providing extensive metadata alongside speech recordings. It is a comprehensive collection of speech recordings from 345 Indian speakers (219 males and 126 females) reading English sentences. The speakers' first languages are from five major Indian languages: Hindi, Kannada, Malayalam, Tamil, and Telugu. Each participant contributed approximately 40 sentences, resulting in a total of 32.03 hours of non-native accented English speech. The data was collected in controlled environments such as classrooms and seminar halls to minimize background noise and ensure high audio quality. High-quality microphones (Scarlett Solo Studio CM25) were used, with recordings sampled at 44.1 kHz and 16-bit resolution, ensuring consistent and clear audio. In addition to the speech data, the NISP dataset includes extensive metadata, such as the speakers' native language (L1), language used during schooling, second language (L2), geographic location, and physical characteristics (age, gender, height, shoulder size, and weight). This rich metadata supports various speaker profiling applications and enhances the dataset's utility for research in speaker recognition and accent identification. The diverse linguistic and accent data, combined with the detailed metadata, make the NISP dataset an ideal resource for training accent conversion models.

As the cascade model requires parallel data for training, it was essential to generate native-like

reference speech for each utterance in the NISP dataset. To achieve this, Google Text-to-Speech (TTS) was utilized to create parallel speech data. The transcript of each utterance from the NISP corpus was synthesized using Google TTS in a female voice with a North American accent.

3.3 Adaptation of the Cascade Model for Language Assessment

In adapting the original 2-stage model proposed by W. Huang et al. (2021) for the accent conversion task, I have made several key adjustments. The focus here is on correcting the accent and altering the voice within a single stage, which aligns with the goal of enhancing fairness for language assessments. Only the first stage of the original model, which is a transformer-based seq2seq model, will be trained. This model will learn to map the accented L2 speaker's mel spectrogram to a native spectrogram.

In Huang's work, a neural vocoder based on Parallel WaveGAN (PWG) was used to improve the naturalness of the converted speech. Although the neural vocoder outperformed the phase reconstruction vocoder, the focus of this study is on accent conversion. For this reason, a Griffin-Lim vocoder will be used in this study to read the mel-spectrogram and generate the converted speech.

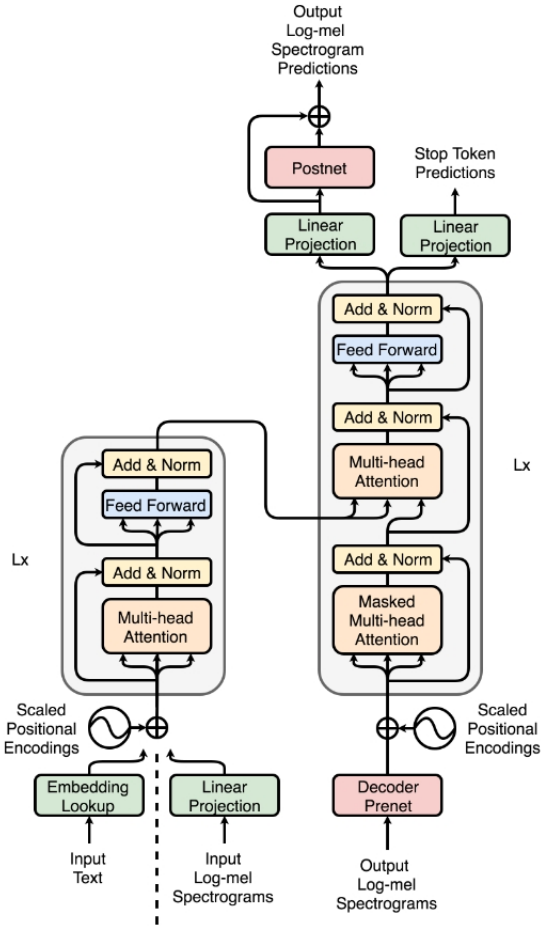


Figure 5: VTN model structure (W. Huang et al., 2020)

The transformer-based sequence-to-sequence (seq2seq) model utilized in this study is based on the earlier work by W. Huang et al., known as the Voice Transformer Network (VTN) (W. Huang et al., 2020). This model employs a pretraining technique to transfer knowledge from text-to-speech (TTS) models, which leverage large-scale, easily accessible TTS corpora to convert speech from a source to a target speaker while preserving the linguistic content. The VTN models, initialized with these pretrained parameters, are capable of generating effective hidden representations, resulting in high-fidelity, highly intelligible converted speech.

I followed Huang's method(Figure 6) in building the VTN model. The pretraining process begins with the decoder, which is trained using a large-scale TTS corpus to develop a conventional TTS model. This training ensures the decoder is well-equipped to generate high-quality speech with accurate hidden representations. Subsequently, the encoder is pretrained in an autoencoder style, where the pretrained decoder remains fixed. This process enables the encoder to encode input speech into hidden representations that the decoder can recognize, thereby facilitating the generation of high-quality converted speech.

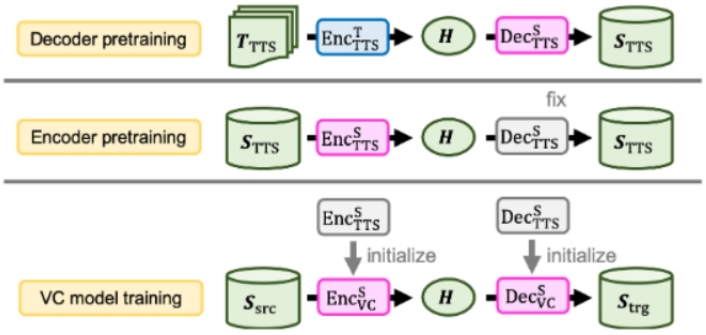


Figure 6: Illustration of TTS pretraining for VTN (W. Huang et al., 2020)

The VTN Model adopted in this study can be summarized as below,

Encoder Stack:

Input:	Log-mel spectrograms
Attention Dimension:	384
Attention Heads:	4
Number of Encoder Blocks:	6
Linear Units:	1536

Decoder Stack:

Output:	Log-mel spectrograms
Attention Dimension:	384
Attention Heads:	4
Number of Decoder Blocks:	6
Linear Units:	1536
Decoder Prenet Layers:	2
Decoder Prenet Units:	256
Postnet Layers:	5
Postnet Filters:	5

This configuration reflects the essential parameters of the adapted model used for the accent conversion task. The focus on using a Transformer-based seq2seq model and a Griffin-Lim vocoder helps maintain a simple yet effective structure for the task at hand.

3.4 Evaluation - Word Error Rate and Mean Opinion Score

To assess the effectiveness of the Foreign Accent Conversion (FAC) model, both objective and subjective evaluation methods will be employed. This section outlines the procedures for these evaluations, including the use of an Automatic Speech Recognition (ASR) system to measure word error rate (WER) as well as character error rate (CER), and conducting a Mean Opinion Score (MOS) test to gather subjective ratings from human listeners.

- The first evaluation method involves measuring the Word Error Rate (WER) and character error rate (CER) of the speech recognition results. This metric quantifies the accuracy of speech recognition by comparing the recognized text with the reference text. An English ASR model, trained exclusively on standard English datasets, will be used for this evaluation. The model will not have been exposed to non-native accents, ensuring that it assesses the speech based purely on its standard English training. The ASR model will be used to transcribe the original non-native speech samples from the L2-ARCTIC and NISP datasets as well as the converted speech samples produced by the proposed FAC model. The WER and CER will be calculated for both sets of transcriptions by comparing them to the reference transcriptions provided in by the datasets. A decrease in WER or CER for the converted speech compared to the original non-native speech will indicate an improvement in the speech's intelligibility and alignment with standard English pronunciation.

- The second evaluation method involves conducting a Mean Opinion Score (MOS) test to gather subjective ratings from human listeners. This test will assess perceived language proficiency, comprehensibility, the strength of the non-native accent in the speech samples, and the similarity of speaker identity. A diverse group of participants will be selected to rate the speech samples. Participants will choose from five options based on their perceptions, with each option corresponding to a score from 1 to 5. For language proficiency, 1 represents the weakest proficiency and 5 represents the strongest proficiency. Similarly, for the perceived strength of the non-native accent, 1 represents the weakest accent and 5 represents the strongest accent.

Participants will be presented with both the original non-native speech samples and the converted speech samples. The MOS ratings will be averaged for each speech group and each criterion. Higher MOS ratings for the converted speech in terms of language proficiency and lower ratings in terms of accent strength will indicate the effectiveness of the FAC model in neutralizing accents and improving perceived language proficiency.

Together, these evaluations will validate the hypothesis that the FAC model can effectively neutralize accents in English speech, enhancing the fairness and objectivity of language proficiency assessments.

3.5 Ethical considerations

Given the nature of the study, it is important to ensure that ethical standards are maintained, particularly in relation to data usage, potential biases, and the transparency of research findings.

- For this research, I have utilized the CMU-ARCTIC, L2-ARCTIC, and NISP corpora—collections of non-native English speech that are open-source and freely available. To enrich our cascade model with parallel speech data, I integrated Google's Text-to-Speech (TTS) service. It's important to highlight that no new data was gathered from individuals for this study. I relied on these pre-recorded and synthesized datasets, which means there was no need for direct human participation in the data collection process. All contributors to the projects from which I sourced the data were fully informed about how their contributions would be used. Participation was entirely voluntary, and the data was vetted by respected institutions to ensure its quality and reliability. These include the Language Technologies Institute at Carnegie Mellon University, the Perception, Sensing, and Instrumentation Lab at Texas A&M University, the National Institute of Technology in Karnataka, India, and the Learning and Extraction of Acoustic Patterns (LEAP) lab at the Indian Institute of Science in Bangalore. The L2-ARCTIC corpus is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), and the NISP dataset is available under the Creative Commons Attribution 4.0 International License. These licenses permit the datasets to be used freely for research without any legal or ethical concerns.

- Potential Bias and Mitigation: Despite its comprehensiveness, the NISP dataset may exhibit biases due to regional and dialectal variations within the native languages, which can influence the English accent. Additionally, while the dataset aims to cover a wide array of speakers, specific speaker characteristics like native language and geographic origin could introduce variability. To mitigate these biases and control variations in accent, this study focuses exclusively on utterances from speakers whose native language is Hindi. It is essential to ensure the data used for model training is representative and focused. Consequently, the average quality of each speaker's recordings was assessed, and those with significant errors (e.g., mispronunciations, long pauses, or reading mistakes) were excluded. This process resulted in the selection of 103 speakers and 3,869 utterances, ensuring that the training data consisted of high-quality speech samples. This approach helps maintain the integrity and reliability of the accent conversion models developed using this dataset.

- Transparency and Replicability: To ensure the transparency and replicability of the research, the following steps have been taken:

- **Open Access to Code and Models:** The code used in this research is available via GitHub[1], allowing other researchers to review, replicate, and build upon the work. Fine-tuned models are available on the Hugging Face Hub, providing access to the specific implementations used in this study. URLs to each model are provided in the relevant sections of the thesis.
- **Detailed Experimental Procedures:** All steps and details necessary to reproduce the experiments are thoroughly documented in the methodology and experiments section of the thesis. The dataset is publicly available for download and use, ensuring that others can conduct similar experiments and validate the findings. The generated parallel data used in this experiment is also provided on hugging face.
- **Consideration of Hardware and Randomness:** The outcomes of the experiments may vary slightly due to elements of randomness inherent in model training and differences in hardware used. Experiments were conducted on the University of Groningen's high-performance cluster, Habrók, which may influence performance. This is noted to provide context for any potential variability in results.

By addressing these ethical considerations, this research aims to maintain high standards of integrity, transparency, and accountability, ensuring that the findings contribute responsibly to the advancement of knowledge in the field of accent conversion and language assessment.

With this, I wrap up the methodology section, which has given a broad overview of the strategies I have adopted in this study. Moving forward, the next section will delve into the experimental setup, providing detailed insights into the datasets employed and the specific parameters of the proposed models.

[1] https://github.com/Jasmijn888/vt_fac

4 Experimental Setup

This chapter outlines the detailed setup of the experiments conducted to evaluate the effectiveness of the speaker-independent Foreign Accent Conversion (FAC) model. The setup includes comprehensive descriptions of data preparation, model training plans, hyperparameters used, evaluation methods, and the software and hardware configurations employed.

4.1 Data Splitting of Subsets

In the L2-ARCTIC dataset, the subdataset SVBI, which consists of recordings from a female native Hindi speaker, is chosen for this study. There are 1,132 utterances in this subdataset. Meanwhile, the BDL subset(male, native American) from CMU-ARCTIC is used as the target speaker. Based on the audio file IDs, sorted in descending order, the first 1,000 utterances are used as the training set, the next 50 as the development set, and the final 50 as the evaluation set.

In the NISP dataset, there are a total of 103 speakers with Hindi as their native language. Each speaker has provided approximately 40 recordings, each about 10 seconds in length. Among these recordings, the first 6 recordings of each speaker contain identical content, while the rest have different content. The purpose of this experiment is to train a speaker-independent accent conversion model capable of adapting to various speakers.

The data is divided as follows:

- Test Set for Unseen Speakers: Recordings from 3 speakers (speaker IDs 0101-0103) are excluded from the training set to test the model's ability to convert accents for unseen speakers.
- Validation and Evaluation Sets: From the remaining 100 speakers (speaker IDs 0001-0100), the first 6 identical content recordings are separated. The recordings from the first 50 speakers (speaker IDs 0001-0050) serve as the validation set, and those from the last 50 speakers (speaker IDs 0051-0100) serve as the evaluation set.
- Training Set: All remaining non-repetitive recordings from the 100 speakers constitute the training set.

Ultimately, there are 300 recordings each in the validation and evaluation sets, while the training set contains 3,960 recordings. This design ensures that the model is trained on recordings with Hindi accents from different speakers, enabling it to adapt to various speakers and become a speaker-independent accent conversion model. During the inference phase, the model is expected to effectively convert accents for trained as well as unseen speakers.

This data split strategy is effective for several reasons:

- Representation of Variability: By including recordings from a wide range of speakers in the training set, the model can learn the diverse characteristics of Hindi-accented English. This variability is crucial for developing a speaker-independent model.

- Validation and Evaluation with Identical Content: Using identical content recordings for validation and evaluation ensures consistency in assessment and helps in accurately measuring the model's performance on familiar tasks. This also allows for easier comparison between the validation and evaluation stages.
- Generalization to Unseen Speakers: By excluding recordings from specific speakers (IDs 0101-0103) entirely from the training process, the model's ability to generalize to new, unseen speakers can be tested. This aspect is vital for assessing the practical applicability of the model in real-world scenarios.
- Focused Training on Non-repetitive Content: The training set consists of non-repetitive recordings, which ensures that the model is exposed to varied linguistic content. This diversity helps the model to learn broader linguistic patterns rather than overfitting to specific repeated phrases.

This data split approach is designed to maximize the model's ability to generalize across different speakers and to ensure robust performance.

4.2 Experiments

In this research, I have designed and conducted four experiments to develop a speaker-independent sequence-to-sequence foreign accent conversion model. These experiments incrementally explore and refine the model with the goal of achieving effective accent conversion in oral examinations, thereby enhancing the fairness of these assessments. A detailed introduction to each experiment is presented in the following subsections.

4.2.1 Experiment 1: One-Stage Speaker-Dependent Accent Conversion - Baseline Model Experiment

The first experiment aims to adjust and evaluate a speaker-dependent model to verify its feasibility in converting accents and voices for oral examinations, serving as a baseline for subsequent experiments.

Initially, the speaker-dependent model proposed by W. Huang et al. (2021) is applied, executing only the first step of the model. According to W. Huang et al.'s theory, this step should generate speech with changes in both voice and accent. The model is initialized with pretrained TTS model, which is trained on M-AILABS dataset, using the speech recorded by an American female. In total 15,200 utterance, about 32 hours speech recording is used for taining. Model structure as provided in Chapter 3, the hyperparameters used in training are as below:

Taining Steps:	50,000
Learning Rate:	0.00008
Batch Size:	16
FFT Size:	1024
Hop Size:	256
Window Type:	Hanning
Gradient Norm:1.0	1.0
Sampling Rate:	16,000

The experiment utilizes the L2-ARCTIC(Zhao et al., 2018) and CMU-ARCTIC(Kominek & Black, 2004) datasets for training and testing. Specifically:

- Source Speaker: L2 speaker SVBI, a female Indian with Hindi as her native language.
- Target Speaker: BDL, a male native American speaker.

The experimental results are evaluated using Word Error Rate(WER), Character Error Rate(CER) and Mean Opinion Score(MOS). The experiment compares the WER and CER of the original and accent-altered audio to assess whether there is an improvement, where an Automatic Speech Recognition (ASR) system[2] trained solely on native accents LJ-speech dataset is used. Participants of MOS survey are asked to rate the audios, focusing on their ability to perceive changes in accents and scoring the speakers' English proficiency based on their understanding of both the original and altered audios.

For the MOS test, 5 audio samples that have undergone accent correction were randomly selected from the evaluation set, along with their corresponding original audios. Subsequently, 20 participants were invited to assess these recordings and answer the following question:

- Based on this recording, how would you rate the speaker's proficiency in English?
  - a. Very poor
  - b. Poor
  - c. Average
  - d. Good
  - e. Excellent
- To what extent are you able to understand the speaker?
  - a. Do not understand at all
  - b. Slightly understand
  - c. Partially understand
  - d. Mostly understand

[2]https://huggingface.co/facebook/wav2vec2-base-960h

  - e. Completely understand

- If using North American accent as the standard, do you find the accent in this recording obvious?
  - a. No noticeable foreign accent
  - b. Slight foreign accent
  - c. Neutral accent, neither very noticeable nor obscure
  - d. Somewhat noticeable foreign accent, but not severe
  - e. Very strong foreign accent
- To what extent do you believe that this recording and the reference recording are from the same speaker? (Provide another original recording from the speaker)
  - a. Not at all the same speaker
  - b. Mostly different, unlikely the same speaker
  - c. Somewhat similar, likely the same speaker
  - d. Mostly the same, very likely the same speaker
  - e. Identical, definitely the same speaker

Candidates for this study are selected based on the following requirements:

- Educational Attainment: Participants must possess a bachelor's degree or higher, indicating a solid foundation of academic knowledge and critical thinking skills.
- English Proficiency: A minimum English proficiency level of CEFR C1 is mandatory. This ensures that participants have an advanced command of the English language, capable of expressing ideas fluently and spontaneously without strain.
- The study includes two groups of participants: one group of 10 individuals who self-report significant exposure to the Indian accent and demonstrate a clear understanding of its nuances. The other group consists of 15 participants who self-report limited familiarity or experience difficulty understanding Indian-accented English. This distinction is crucial to ensure a diverse sample capable of accurately assessing the impact of accents on language evaluations and to facilitate comparative analysis.

The purpose of this experiment is threefold: to verify whether the model can achieve the results described by the author, to demonstrate whether changing the accent and voice can improve the fairness of examinations, and to provide a baseline model for further experiments.

4.2.2 Experiment 2: Training on Various Speakers Dataset Initializing by Checkpoints of Pretrained TTS Model

This experiment aims to train the accent conversion model on the diverse NISP dataset, initializing the training with checkpoints from a pretrained Text-to-Speech (TTS) model. The dataset splitting methodology is detailed in the Chapter 3. The goal of this experiment is to assess whether training on a diverse dataset of various speakers improves the generalization capabilities of the accent conversion model. Specifically, this experiment evaluates if the model can generalize to both the speakers included in the training data and to unseen speakers. Initial evaluation will be conducted using WER and CER. If the results are satisfactory, MOS evaluations will be implemented. The hyperparameters used are the same as in Experiment 1(4.2.1).

4.2.3 Experiment 3: training on various speakers dataset initializing by checkpoints of Experiment 1

This experiment aims to train the accent conversion model on the diverse NISP dataset, initializing the training with checkpoints from Experiment 1. The dataset splitting methodology is detailed in the Chapter 3. The goal of this experiment is to assess whether training on a diverse dataset of various speakers improves the generalization capabilities of the accent conversion model. Specifically, this experiment evaluates if the model can generalize to both the speakers included in the training data and to unseen speakers. Additionally, it examines whether using checkpoints from a speaker-dependent model, which has already been trained on the same accent, improves performance compared to Experiment 2. The hyperparameters used are the same as in Experiment 1(4.2.1). Initial evaluation will be conducted using objective metrics such as Word Error Rate (WER) and Character Error Rate (CER). If the results are satisfactory, Mean Opinion Score (MOS) evaluations will be implemented.

4.2.4 Experiment 4: Training on Various Speakers Dataset with Speaker Embedding as Extra Input

This experiment aimed to explore whether incorporating speaker embeddings as an additional input could improve the model's performance. The speaker embeddings were intended to help the model distinguish between content information and speaker identity, potentially enhancing the effectiveness of the accent conversion. Using the same dataset and trained from scratch the model was trained with speaker embeddings added as extra input features.

To ensure thorough training, the model was trained for 200,000 steps, compared to previous experiments. The increased number of training steps was necessary because this experiment did not initialize from any prior checkpoints. The effectiveness of the model was determined using WER and CER. If the results are satisfactory, MOS evaluations will be implemented.

For the speaker embeddings, I adapted the d-vector approach proposed by Wan, Wang, Papir, and Moreno (2017), which is developed for speaker verification. Each speaker's embedding was calculated using all their utterances, and the average embedding was used to represent the speaker, as shown in Figure 7 . This method aimed to provide a robust speaker identity feature that could aid in accent conversion.

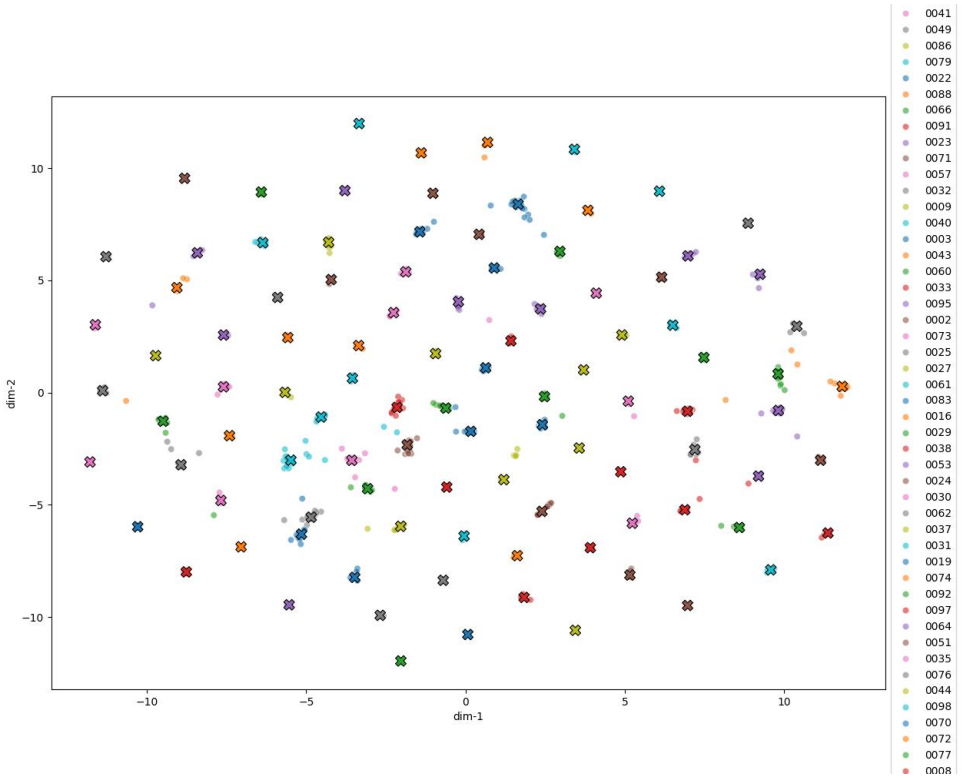


Figure 7: D-vector Speaker Embeddings(Wan et al., 2017)

The speaker embedding proposed by Wan et al. is 256 dimensions. To balance between the mel-spectrogram features and speaker embeddings, the 256-dimensional embedding is reduced to 64 dimensions using Principal Component Analysis (PCA). This 64-dimensional speaker embedding is then replicated to match the length of the speech signal and concatenated to each frame's 80-dimensional mel features.

Besides the change in the training data from 80 dimensions to 144 dimensions, the rest of the hyper-parameters remain the same as in Experiment 1(4.2.1).

5 Results

This chapter presents the results of the experiments conducted to evaluate the effectiveness of the Foreign Accent Conversion (FAC) models proposed in section. Each experiment's outcomes are detailed, focusing on key metrics such as Word Error Rate (WER), Character Error Rate (CER), and Mean Opinion Score (MOS). Figures and tables summarizing the data will be provided to illustrate the findings.

5.1 Performance Comparison of Different Experiments

Table 2: Performance Comparison of Different Experiments

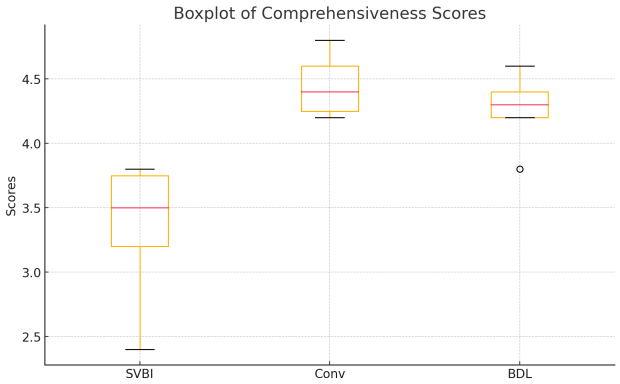
Method	**CER/WER(0%-100%)**	**Proficiency(1-5)**	**Comprehensibility(1-5)**	**Accentedness (1-5)**	**Similarity**
Source (SVBI)	7.4/13.3	F:2.96	3.76	4.56	4.44
		NF:3.08	3.4	4.56	4.42
Target (BDL)	4.5/6.3	F:4.36	3.92	1.32	-
		NF:4.72	4.68	1.2	-
Experiment1	25.4/45.6	F:4.22	3.92	2.16	1.4
		NF:3.8	4.44	2.68	1.2
Experiment2	39.9/78.2	-	-	-	-
Experiment3	45.1/86.5	-	-	-	-
Experiment4	-/-	-	-	-	-

Note:

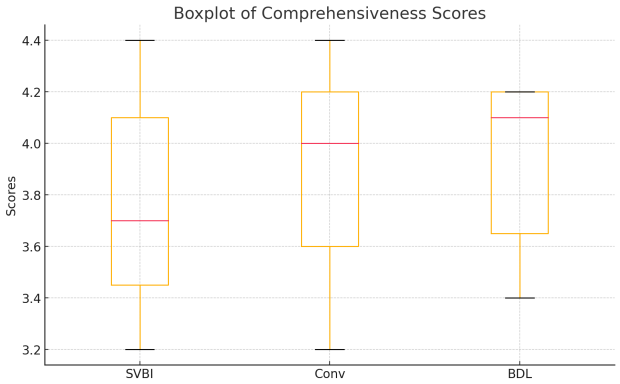
- F refers to the results from participants who claim to be familiar with the Indian accent, and NF refers to those who are not.
- The speech generated from Experiment 4 is predominantly characterized by a significant amount of noise and muffled sounds, resulting in extremely poor intelligibility. Consequently, the Automatic Speech Recognition (ASR) system was unable to produce a valid transcription, and therefore, neither the Character Error Rate (CER) nor the Word Error Rate (WER) could be calculated.

5.2 Results of MOS

The MOS results of Experiment 1 can be illustrated as follows:

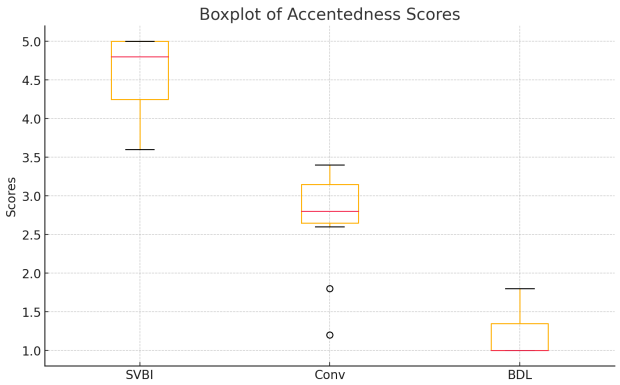


(a) Comprehensibility Scores by Participants Not Familiar with Indian Accents

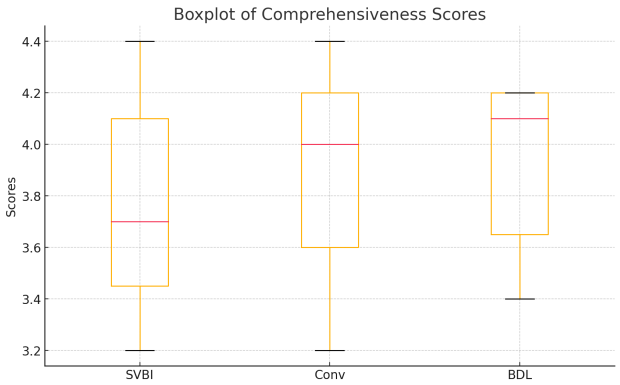


(b) Comprehensibility Scores by Participants Familiar with Indian Accents

Figure 8: Comprehensibility Scores by Different Participant Groups

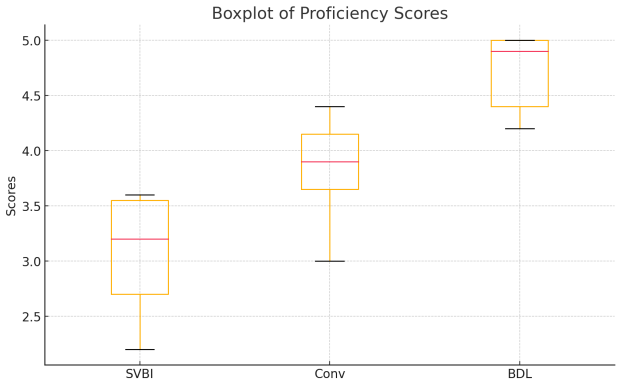


(a) Accentedness Scores by Participants Not Familiar with Indian Accents

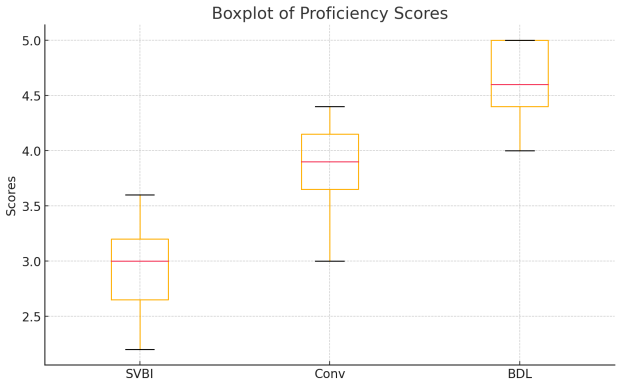


(b) Accentedness Scores by Participants Familiar with Indian Accents

Figure 9: Accentedness Scores by Different Participant Groups

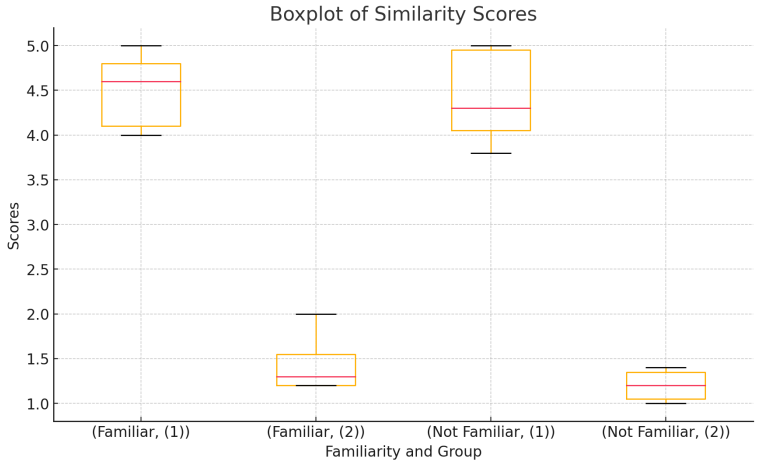


(a) Proficiency Scores by Participants Not Familiar with Indian Accents



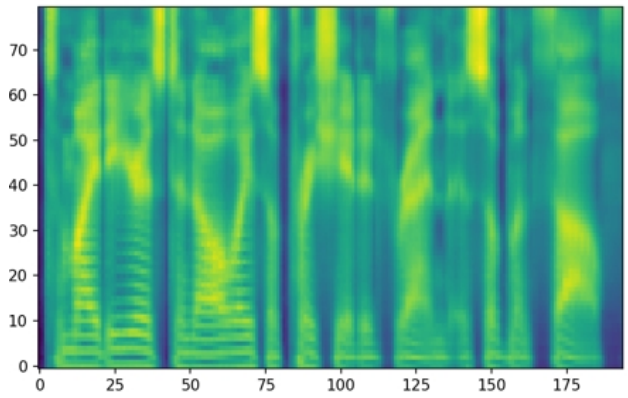
(b) Proficiency Scores by Participants Familiar with Indian Accents

Figure 10: Proficiency Scores by Different Participant Groups

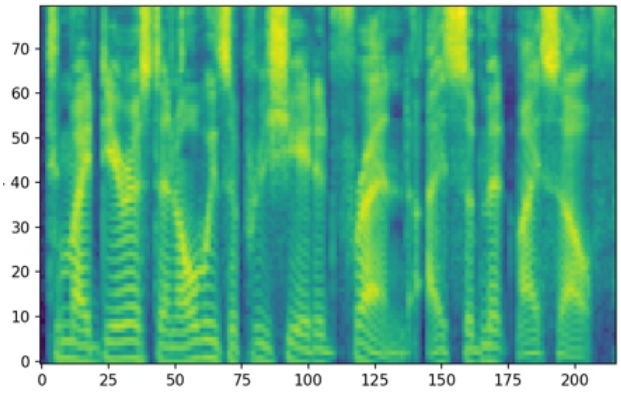


(a) Similarity Scores by Different Participant Groups

5.3 Results of Intermediate Training Progress

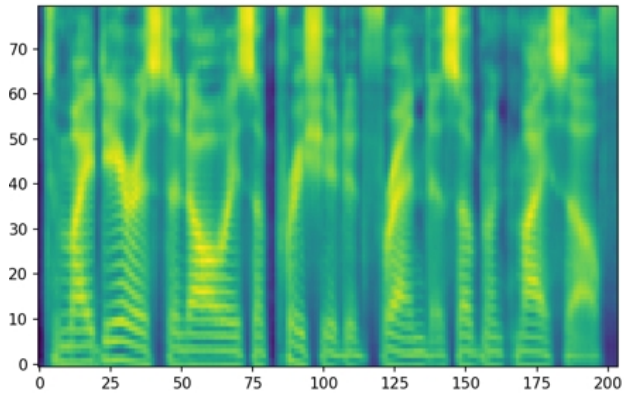


(a) Melspectrogram of Converted Speech at 5,000 Steps in Experiment 1

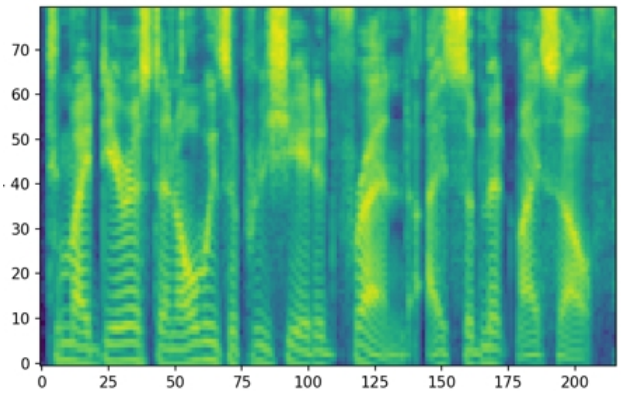


(b) Melspectrogram of Target Speech

Figure 12: Comparison of Melspectrograms in Experiment 1 at 5,000 training steps

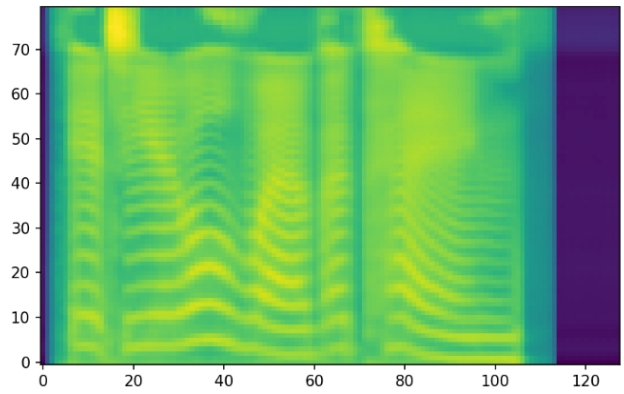


(a) Melspectrogram of Converted Speech at 50,000 Steps in Experiment1

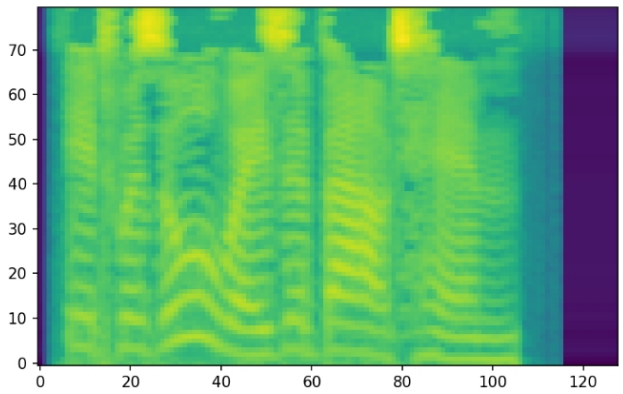


(b) Melspectrogram of Target Speech

Figure 13: Comparison of Melspectrograms in Experiment 1 at 50,000 training steps

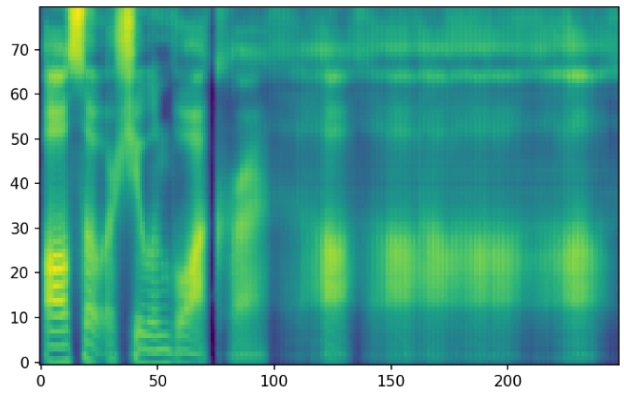


(a) Melspectrogram of Converted Speech at 50,000 Steps in Experiment 2

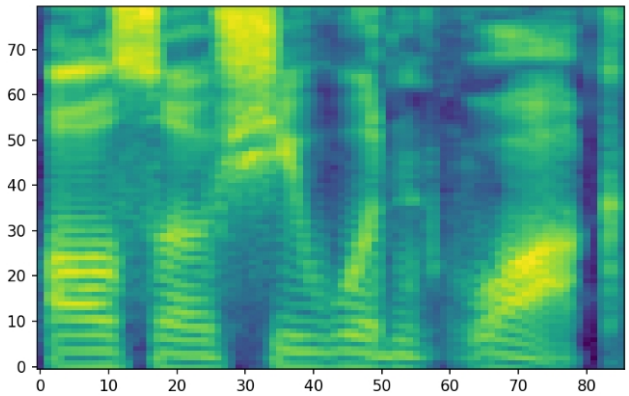


(b) Melspectrogram of Target Speech

Figure 14: Comparison of Melspectrograms in Experiment 2 at 50,000 training steps

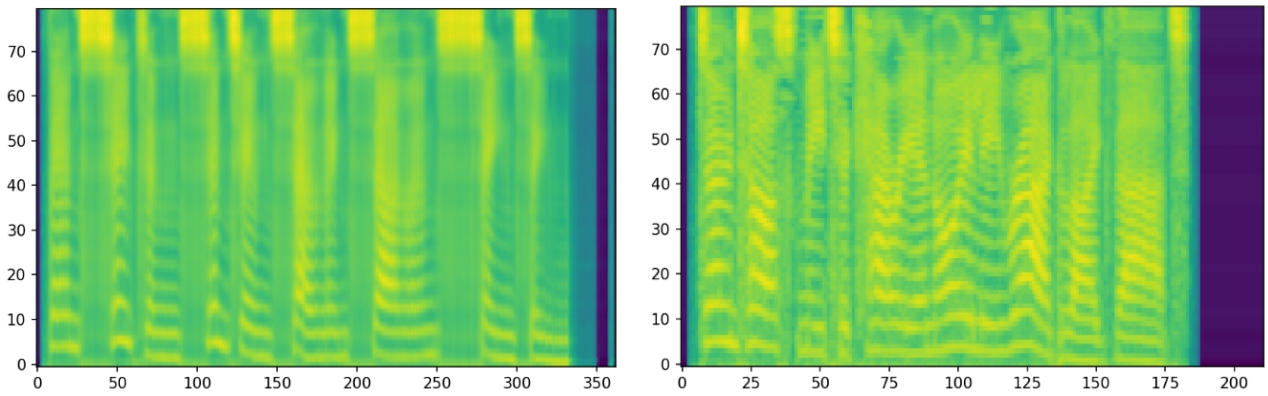


(a) Melspectrogram of Converted Speech at 50,000 Steps in Experiment 3



(b) Melspectrogram of Target Speech

Figure 15: Comparison of Melspectrograms in Experiment 3 at 50,000 training steps



(a) Melspectrogram of Converted Speech at 200,000 Steps in Experiment 4

(b) Melspectrogram of Target Speech

Figure 16: Comparison of Melspectrograms in Experiment 4 at 200,000 training steps

6 Discussion

This chapter presents the results of the experiments conducted to evaluate the effectiveness of the speaker-independent Foreign Accent Conversion (FAC) model. The evaluation focuses on validating two hypotheses: First, that an Accent Conversion Model could bring necessary improvements to oral examinations, enhancing the fairness and objectivity of assessments; and second, that developing a Speaker-Independent model that could be widely applied is feasible. This section will delve into the discussion of the experimental results to assess the validity of the two hypotheses and explore the potential implications of my findings for language assessment practices.

6.1 Validation of the First Hypothesis

6.1.1 Overview of the Findings

My initial hypothesis suggested that integrating an Accent Conversion Model into oral examinations is crucial for reducing the influence of non-native accents on language proficiency assessments. The outcomes of my experiments, especially the first one, have substantiated this hypothesis. By applying a speaker-dependent model to transform accented speech, the modified speech exhibited a diminished accentedness, which corresponded to higher proficiency scores for the speakers and an increased level of speaker anonymity. These improvements were evidenced by the Mean Opinion Score (MOS) evaluations, which reflected a more favorable perception of the language proficiency and a weaker accentedness in the converted speech samples.

6.1.2 Analysis of Results

The experiments conducted have yielded intriguing and somewhat controversial results, particularly when juxtaposing subjective evaluations against objective measures.

6.1.3 Subjective Evaluations

The Mean Opinion Score (MOS) evaluations indicate an enhanced perception of language proficiency. For participants who claim to be familiar with the Indian accent, the ratings improved from 2.96 to 4.22, and for participants who are not familiar with the Indian accent, the ratings improved from 3.08 to 3.8, representing an increase of 42% and 23%, respectively. This reflects a more positive assessment of the speakers' English proficiency when hindi accent is neutralized.

Additionally, the perceived accentedness was significantly reduced. For participants familiar with the Indian accent, the accentedness score dropped from 4.56 to 2.16, and for those not familiar with the Indian accent, it dropped from 4.56 to 2.68, representing decreases of 53% and 41%, respectively. This indicates that the converted speech was perceived as less accented.

Moreover, the anonymity of the speakers was enhanced, as evidenced by the similarity rating to the original speaker's voice dropping significantly. For participants familiar with the Indian accent, the similarity rating dropped from 4.44 to 1.04, and for those not familiar with the Indian accent, it dropped from 4.42 to 1.2, representing decreases of 76% and 72%, respectively. This suggests that the converted speech samples were less identifiable in terms of the original speaker's unique

characteristics.

Furthermore, the comprehensibility of the speech was also improved. The comprehensibility score increased from 3.76 to 3.92 for participants who are familiar with the Indian accent, and from 3.4 to 4.44 for those who are not familiar with the Indian accent. This represents increases of 4% and 31%, respectively. This demonstrates that the converted speech was easier to understand. especially for listeners who are less familiar with the accents.

6.1.4 Objective Evaluations

The ASR outcomes were contrary to expectations. Specifically, the Character Error Rate (CER) saw an increase from 7.4% to 25.4%, while the Word Error Rate (WER) experienced a significant rise, from 13.3% to 45.6%. Examples illustrating these ASR results can be found in the Appendix, in Table 5.

Based on the analysis of the transcriptions, the ASR results from the converted speech exhibit several discrepancies compared to the original speech's ASR results. These issues include the merging of words incorrectly, the insertion of extraneous words and the errors in phoneme recognition. For example:

The original phrase "NOT A WHEEL MOVED IN HIS EMPIRE" was incorrectly transcribed as "NOT AVEALD MORE THAN HIS EMPIRE FIRE" during the ASR process. This error includes a word merging mistake where "A WHEEL" was recognized as "AVEALD", and an additional word "FIRE" was erroneously appended at the end of the sentence. In the phrase 'HE HAD BECOME A MAN VERY EARLY IN LIFE,' the pronunciation of 'VERY' in the converted speech was misinterpreted by the ASR system, resulting in a phoneme recognition error where 'VERY' was incorrectly transcribed as 'WEARY'.

Based on these observations, my analysis is as follows:

- The converted speech may have more uniform and mechanical pauses, lacking the natural prosody and rhythm of human speech. This can cause the ASR system to incorrectly merge two separate words into one, as the natural variations in pauses and rhythm are crucial cues for speech segmentation and word recognition. For example, "A WHEEL" being recognized as "AVEALD" could be due to the lack of natural pauses.

- Phoneme recognition errors indicate that the converted pronunciation might be similar to but not exactly matching the standard English pronunciation. This could be due to the conversion model failing to capture the subtle nuances of the target accent, leading to phonemes being misrecognized by the ASR system. For instance, "VERY" being recognized as "WEARY" suggests that the converted speech has ambiguities or inaccuracies in its phonetic details.

- The issue of adding extra words at the end of sentences may be related to the Transformer model used. The attention mechanism in Transformer models can struggle with long-range dependencies, particularly towards the end of generated sequences. The attention mechanism might fail to focus correctly when processing end-of-sequence information, leading to the

generation of extraneous words. This could also be influenced by the quality and quantity of the training data, replicating similar errors present in the data.

Based on the results of Experiment 1, the Mean Opinion Score (MOS) indicated the positive effects of the Foreign Accent Conversion (FAC) model in improving the naturalness of speech. However, contrary to expectations, the results from the Automatic Speech Recognition (ASR) showed that the word error rate (WER) actually increased for the accent-modified audio during recognition. This outcome does not align with my previous assumption that accent conversion would decrease the recognition error rate.

To delve deeper into this phenomenon, I reviewed previous studies which discussed the feasibility of using objective measures to predict subjective outcomes in voice conversion research. Some prior works on FAC used the Character Error Rate(CER) or Word Error Rate(WER) as indirect measures of accentedness, with the expectation that reducing accentedness could also lead to a reduction in error rates (Das, Kinnunen, Huang, et al., 2020)(W. C. Huang et al., 2022). Furthermore, according to the conclusions drawn in "Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion(W. Huang & Toda, 2023)", the authors calculated the linear correlation coefficients between accentedness and CER/WER through several experiments with different accent modifications. The findings indicated a weak and insignificant correlation between accentedness and CER/WER.

Thus, I conclude that when considering accentedness, there are factors at play beyond intelligibility. Therefore, using CER/WER as an objective measure for FAC is unreliable. However, observing its transcription results can assist in better understanding the model's performance. On the contrary, the results from the Mean Opinion Score (MOS) will be highly valued as they provide a subjective assessment of speech quality that complements the objective metrics by reflecting human listeners' perceptions.

6.1.5 Implications for Language Assessments

The success of the Accent Conversion Model in reducing accentedness and enhancing speech intelligibility, points to its potential as a valuable tool in the realm of language proficiency evaluations. By mitigating the influence of non-native accents, the model has the capacity to democratize the assessment process, ensuring that evaluations are more closely aligned with the actual linguistic competence of the test-takers.

The implications of this research are particularly salient in high-stakes language testing scenarios, such as those encountered in educational advancement, professional certification, and immigration processes. In these contexts, the accent of a speaker should not be a decisive factor in determining their language proficiency. The Accent Conversion Model offers a promising avenue for reducing the impact of such biases, leading to more equitable scoring and a fairer assessment process.

Furthermore, the enhanced anonymity provided by the Accent Conversion Model could also be beneficial in blind evaluation settings, where the goal is to assess language proficiency without any influence from the speaker's identity or background. This added layer of objectivity could improve the overall reliability and validity of language assessments.

6.1.6 Comparison with Existing Literature

The experimental results offer valuable insights that can be contextualized within the existing literature on accent familiarity and language assessment. Consistent with previous research highlighting the impact of accent familiarity on comprehensibility (Ockey & French, 2016), the participants in this study who self-reported familiarity with the Indian accent (F) tended to score the original audio from the L2-ARCTIC(SVBI) dataset higher on comprehensibility (3.76 and 3.4 respectively). This suggests that familiarity does play a role in the ability to understand accented speech.

My findings also resonate with the study by B. H. Huang et al. (2016), which discovered that while raters did not significantly differ in their judgments based on their familiarity with an accent, the raters themselves reported a belief that their familiarity could influence their assessments, potentially leading to more lenient evaluations for speakers with familiar accents. This aligns with my experimental observations, where a notable gap was observed in comprehensibility scores between participants familiar with the Indian accent and those who were not. However, when it came to speaker proficiency scores, there was little distinction between the two groups, 2.96 and 3.08 respectively. This suggests that while raters might perceive a familiarity with an accent, this perception does not necessarily manifest in a lenient rating of proficiency.

The experimental results of this study provide further proof that reveals the potential biases of raters towards non-native accents and propose the potential path of using accent conversion technology to improve the fairness of language assessments.

6.2 Validation of the Second Hypothesis

6.2.1 Overview of the Findings

The second hypothesis proposed to develop a speaker-independent, simple-structure machine learning-based model to ensure fair language assessments. Experiments 2, 3, and 4 were designed to test this hypothesis by training the model on various speakers' datasets and initializing with checkpoints from pretrained TTS models or previous experiments or adding speaker embeddings as training feature. Unfortunately, the findings from these experiments indicated that the generated audio quality was poor, with intelligibility being notably inadequate at the sentence level. The WER and CER scores were high, suggesting that the model struggled to produce accurate and intelligible speech.

6.2.2 Analysis of Results

The high WER and CER scores across Experiments 2 and 3 indicate that the model's performance was suboptimal. Despite the use of checkpoints from pretrained TTS models or prior experiments, the model failed to achieve the desired level of accent conversion that would result in fair and unbiased language assessments. The intelligibility issues suggest that the model had difficulty capturing the nuances of different accents while maintaining the clarity and natural flow of speech.

The introduction of speaker embeddings in Experiment 4 as an additional input feature was intended to improve the model's performance by distinguishing between speaker identity and accent characteristics. However, this approach did not yield the anticipated improvements, as the model's

output remained largely unintelligible.

Similar results can also be observed from the comparison of MEL spectrograms. By observing and comparing the MEL spectrograms of the speech generated by the FAC model and the corresponding target speech, it can be found that in the speaker-dependent model, after 5,000 training steps, the Mel spectrogram of the accent-corrected speech already shows some similarities to the target speech. However, the predicted speech length still has a considerable gap compared to the target speech, with the corrected speech being approximately 200 time frames, while the target speech is about 225 time frames(Figure 12). With an increase in training steps to 50,000, the FAC model's generated speech MEL spectrogram becomes closer to the target speech in terms of time frames prediction, both being approximately 215 time steps. The MEL spectrogram details also show significant improvement, especially in the low-frequency range, where the generated speech spectrogram becomes more consistent with the target speech(Figure 13). This indicates that as the number of training steps increases, the FAC model's effectiveness in capturing the features of the target speech improves significantly. The generated speech becomes closer to the target speech in both the length of the time window and the details of the spectrogram, particularly in capturing low-frequency details, where the model performs exceptionally well.

When analyzing the MEL spectrograms generated in speaker-independent experiments, several observations can be made. In Experiment 2, the generated MEL spectrogram is the most similar to the target counterpart. However, the details, especially the pauses between words, are largely incorrect. This suggests that while the overall structure is somewhat similar, the finer details that contribute to natural speech patterns are missing(Figure 14).

In Experiment 3, very few similarities can be observed between the generated and target MEL spectrograms. There is also a significant discrepancy in the number of frames, with the generated MEL spectrogram covering approximately 250 frames compared to the target's 90 frames. This indicates substantial errors in both the duration and the MEL spectral details of the generated speech(Figure 15).

In Experiment 4, there are also few similarities between the generated and target MEL spectrograms. The generated MEL spectrogram is completely incomparable in the high-frequency range, and there is a large difference in the number of frames, with the generated MEL spectrogram having about 350 frames compared to the target's 225 frames. This highlights significant issues in capturing both the frequency details and the temporal structure of the speech. Additionally, the strong intensity observed in the high-frequency range of the MEL spectrogram corresponds with the presence of significant noise in the audio(Figure 16).

6.2.3 Analysis of Failure

- Insufficient Training Data:
  One potential reason could be the insufficient size or diversity of the training data. The inadequacy of the training data can be examined through two critical lenses: the number of speakers included and the overall volume of data.

One significant issue lies in the limited number of speakers used for training the model. With only 100 speakers employed to capture the nuances of Indian accents, the dataset may not be extensive enough to encompass the full range of phonetic and prosodic variations present within this demographic. Accents are highly variable, even within a single linguistic community, and a more extensive and representative sample would be required to ensure that the model learns the necessary patterns to perform effectively across a diverse array of speaking styles.

Additionally, the total amount of training data may be insufficient for the model to achieve robust generalization. In the case of the speaker-dependent model used as a reference, a successful accent conversion model was trained using a combination of 32 hours of pre-trained TTS checkpoint and an additional hour of parallel data. This substantial dataset allowed the model to learn the intricacies of the specific speaker's accent and voice. In contrast, when attempting to generalize to a speaker-independent model, the second and third experiments utilized in total around 11 hours of training data, with each speaker contributing around 7 minutes speech recording. Given the increased complexity of learning to convert accents across multiple speakers without relying on speaker-specific characteristics, this amount of data is likely insufficient for the model to develop the generalizable features needed for effective accent conversion.

The fourth experiment introduced speaker embeddings to the training process, altering the model's architecture and necessitating a training approach that could not leverage the checkpoints from the TTS pre-training phase. This change required the model to learn from scratch to handle richer information, such as the additional speaker identity cues provided by the embeddings. With only 11 hours of data, the model was undertrained, which is a likely explanation for its inability to achieve the desired performance.

- Overfitting to Training Data:
  Overfitting occurs when a model learns the training data too well, and prevents the model from generalizing to new, unseen data. This phenomenon can manifest as high error rates and subpar performance during evaluation.

  A comparative analysis of Experiment 2 and Experiment 3 reveals a critical difference in their approaches and outcomes. Experiment 3 was built upon the checkpoint from Experiment 1, which was a speaker-dependent model trained on a specific Indian speaker's data for one hour. This checkpoint was then used to continue training for an additional six hours, which may have led the model to deeply ingrain the characteristics of that particular speaker.

  In contrast, Experiment 2 initiated training with a checkpoint from a pretrained TTS model, which had a broader exposure to various speakers and speech patterns due to its initial stage training. Despite the fact that both experiments were eventually trained on a diverse dataset of 100 speakers, the foundation and the initial bias introduced by the checkpoints could have significantly influenced the model's performance.

  Experiment 3's worse performance compared to Experiment 2 could be attributed to the data

imbalance caused by the extended training on the specific Indian speaker. After 50,000 steps of training on this particular speaker, the model might have become overly specialized to the nuances of this single speaker's accent and speech patterns. When introduced to the new set of 100 speakers, each with only about seven minutes of training data, the model struggled to generalize and adapt to the broader range of accents.

This imbalance resulted in a training scenario where the model had insufficient data to learn from the diverse speakers and over-relied on the initial one-hour training data from the specific Indian speaker. The lack of exposure to a balanced and varied dataset hindered the model's ability to create a generalized representation of the accents, leading to overfitting.

- Analysis of Model Structure:
  The model's architecture, which employs an attention based a seq2seq framework, is tasked with concurrently converting speaker identity and adjusting the nativeness of the speech. This dual transformation poses a significant challenge, as it requires the model to learn and apply complex mappings that capture the intricate characteristics of different accents and speaker traits.

  The simplicity of the model structure, while potentially beneficial for ease of implementation, may not provide sufficient capacity for learning the complex variations present in natural speech. This limitation becomes particularly evident when compared to more complex pipeline structures used in Foreign Accent Conversion (FAC) models, which break down the conversion process into discrete stages, each addressing a specific aspect of the transformation.

  In contrast to the proposed model, Speaker-Independent Latent Space Conversion (LSC) models employ a more intricate architecture that enables a stepwise approach to accent conversion. Each of these steps is trained on a large dataset independently, allowing for the fine-tuning of each component and the identification of any weak links in the process. This modular approach ensures that the model can be comprehensively evaluated and improved at each stage, leading to a robust and effective accent conversion.

  The model proposed in this study, however, lacks this modular complexity. It attempts to achieve accent conversion and speaker identity transformation within a unified framework without the benefit of separate optimization and assessment of each component. This limitation hinders the model's ability to effectively learn from complex and diverse datasets, particularly when generalizing to new, unseen speakers during inference tasks.

6.2.4 Limitations

While the findings from this study offer promising insights into the potential of Accent Conversion Models for enhancing fairness in language assessments, several limitations must be acknowledged. These limitations highlight areas where the current approach may fall short and suggest directions for future research.

6.2.5 Data Limitations

One of the primary limitations of this study is the size and diversity of the training data. Despite the extensive efforts to collect a comprehensive dataset, the number of speakers and the total duration of the recordings may not be sufficient to capture the full range of phonetic and prosodic variations within the target population. Hindi is the fourth most-spoken first language in the world and serves as an official language in nine states and three union territories, and an additional official language in three other states in India. Native speakers of Hindi are spread over a broad geographical area, resulting in considerable regional variations that affect their English accents. These variations include differences in pronunciation, intonation, and rhythm, which can lead to significant heterogeneity in the data. The lack of control over these regional variations might have introduced high variance in the data distribution, making it challenging for the model to learn consistent speech patterns. Consequently, the model may struggle to generalize across different speakers, reducing its effectiveness in converting accents accurately and consistently.

6.2.6 Evaluation Metrics

Moreover, the MOS ratings, while useful for assessing perceived quality and accentedness, are inherently subjective and may vary depending on the raters' individual biases and perceptions. Additionally, the experimental design could be further improved. Participants were classified into two categories based on their self-claimed familiarity with the Indian accent, without specific quantification of their familiarity level. This approach limits the ability to analyze the relationship between the raters' familiarity with the accent and their scoring. Consequently, the results cannot accurately reflect how varying degrees of familiarity with the Indian accent might influence the evaluation scores.

7 Conclusion

My thesis has explored the potential of accent conversion techniques to enhance fairness in language assessment. The research was driven by the critical need to neutralize accent biases in language proficiency evaluations, which are essential for equitable immigration and integration policies.

7.1 Summary of the Main Contributions

The primary contributions of this research are as follows:

- Test the Speaker-Dependent Model on Indian Accent: The first experiment successfully demonstrated that one-stage model could reduce accentedness, improve speech intelligibility in Indian-accented English, as well as enhance anonymity.
- Speaker-Independent Model Exploration: Experiments 2, 3, and 4 were designed to test the feasibility of a speaker-independent model. Despite the challenges faced, these experiments provided valuable insights into the complexities of generalizing accent conversion across diverse speakers.
- Potential for Fairness in Language Assessments: The findings suggest that accent conversion models could play a significant role in reducing biases in language proficiency evaluations, leading to fairer and more objective assessments.

7.2 Impact and Revelance

The implications of this research extend beyond the academic sphere. By promoting the development of more equitable language assessment tools, this work contributes to the broader goal of inclusivity and diversity in educational and immigration contexts. The accent conversion models have the potential to level the playing field for non-native speakers, ensuring that their language proficiency is assessed based on their actual skills rather than accented speech.

In conclusion, this thesis has demonstrated both the promise and the challenges of using machine learning to neutralize accents in language assessments. While the journey towards a universally fair language assessment tool is ongoing, this research represents a significant step towards that goal. It is hoped that this work will inspire further innovation and contribute to a more equitable evaluation process for speakers of all linguistic backgrounds.

7.3 Future Work

While this research has made significant strides, there are areas that require further exploration:

- Expansion of Datasets: Future models should be trained on larger and more diverse datasets to capture a wider range of accents and speaking styles.
- Enhancement of Model Complexity: More complex models or ensemble approaches could be investigated to better capture the nuances of natural speech.

- Quantification of Familiarity: A more granular assessment of raters' familiarity with accents could provide deeper insights into the impact of accent familiarity on evaluation scores.
- Ethical Guidelines: The development of ethical guidelines for the use of accent conversion technology is necessary to prevent misuse and ensure fairness.

References

Ahola, S., & Halonen, M. (2021). 'broken finnish': Speaker l1 and its recognition affecting rating in national certificates of language proficiency test in finnish. In *Collated papers for the alte 7th international conference, madrid* (pp. 53–57). Retrieved from https://www.alte.org/resources/Documents/ALTE%207th%20International%20Conference%20Madrid%20June%202021.pdf

AlgorithmWatch. (2021). *German federal office for migration and refugees (bamf) and dialect recognition.* Retrieved 2024-05-27, from https://algorithmwatch.org/en/bamf-dialect-recognition/ (Accessed: 2024-05-27)

Aryal, S., & Gutierrez-Osuna, R. (2015). Reduction of nonnative accents through statistical parametric articulatory synthesis. *The Journal of the Acoustical Society of America*, *137*(1), 433–446.

Aryal, S., & Gutierrez-Osuna, R. (2016). Data driven articulatory synthesis with deep neural networks. *Computer Speech & Language*, *36*, 260–273.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. Retrieved from https://doi.org/10.1016/j.cognition.2007.04.005 doi: 10.1016/j.cognition.2007.04.005

Cambridge English. (2013). *Cambridge english: Proficiency speaking sample test with examiner's comments.* (Available at: https://www.cambridgeenglish.org/)

Das, R. K., Kinnunen, T., Huang, W.-C., et al. (2020). Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions. In *Proc. joint workshop for the bc and vcc 2020* (pp. 99–120).

Felps, D., Geng, C., & Gutierrez-Osuna, R. (2012). Foreign accent conversion through concatenative synthesis in the articulatory domain. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(8), 2301–2312.

Huang, B. H., Alegre, A., & Eisenberg, A. R. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41. Retrieved from https://doi.org/10.1080/15434303.2015.1134540 doi: 10.1080/15434303.2015.1134540

Huang, W., Hayashi, T., Wu, Y., Kameoka, H., & Toda, T. (2020). Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. In *Proc. interspeech* (pp. 4676–4680).

Huang, W., Kobayashi, K., & Peng, Y. (2021). A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion. In *Proc. interspeech* (pp. 1329–1333).

Huang, W., & Toda, T. (2023). Evaluating methods for ground-truth-free foreign accent conversion. In *2023 asia pacific signal and information processing association annual summit and conference (apsipa asc).* Retrieved from https://github.com/unilight/seq2seq-vc

Huang, W. C., Cooper, E., Tsao, Y., Wang, H.-M., Toda, T., & Yamagishi, J. (2022). The voicemos challenge 2022. In *Proc. interspeech 2022* (pp. 4536–4540). doi: 10.21437/Interspeech.2022-970

IELTS Asia. (n.d.). *Faq.* Retrieved 2024-05-27, from https://www.ieltsasia.org/ph/faq (Accessed: 2024-05-15)

Kalluri, S. B., Vijayasenan, D., Ganapathy, S., Rajan, R. M., & Krishnan, P. (2020). Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *Proc. interspeech.* (Available at:

https://github.com/iiscleap/NISP-Dataset)

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of l2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, *64*(3), 459–489. Retrieved from https://doi.org/10.3138/cmlr.64.3.459 doi: 10.3138/cmlr.64.3.459

Kerkhoff, A., Poelmans, P., de Jong, J. H., & Lennig, M. (2005). *Verantwoording toets gesproken nederlands.* Retrieved from http://www.cinop.nl (Developed in opdracht van het Ministerie van Justitie van het Koninkrijk der Nederlanden)

Kominek, J., & Black, A. W. (2004). The cmu arctic speech databases. In *Fifth isca workshop on speech synthesis.*

Li, W., Tang, B., Yin, X., Zhao, Y., Li, W., Wang, K., ... Ma, Z. (2020). Improving accent conversion with reference encoder and end-to-end text-to-speech. *arXiv preprint arXiv:2005.09271*. Retrieved from https://doi.org/10.48550/arxiv.2005.09271 doi: 10.48550/arxiv.2005.09271

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for esl assessment. *TESOL Quarterly*, *36*(2), 173–190. Retrieved from https://doi.org/10.2307/3588329 doi: 10.2307/3588329

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of l2 listening comprehension. *Applied Linguistics*, *37*(5), 693–715. Retrieved from https://doi.org/10.1093/applin/amu060 doi: 10.1093/applin/amu060

Quamer, W., Das, A., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2022). Zero-shot foreign accent conversion without a native reference. In *Proc. interspeech* (pp. 4920–4924).

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly*, *40*(4), 715–738. Retrieved from https://doi.org/10.2307/40264305 doi: 10.2307/40264305

Tanaka, K., Kameoka, H., Kaneko, T., & Hojo, N. (2019, May). ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms. In *Proceedings of the ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6805–6809).

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2017). Generalized end-to-end loss for speaker verification. In *Ieee international conference on acoustics, speech and signal processing (icassp).*

Winke, P., & Gass, S. (2012). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *0*(0), 1–24. doi: 10.1002/tesq.73

Xi, X., & Mollaun, P. (2011). Using raters from india to score a large-scale speaking test. *Language Learning*, *61*(4), 1222–1255. doi: 10.1111/j.1467-9922.2011.00667.x

Zang, X., Xie, F., & Weng, F. (2022, November 1). Foreign accent conversion using concentrated attention. In *2022 ieee international conference on knowledge graph (ickg).* IEEE. Retrieved from https://doi.org/10.1109/ickg55886.2022.00056 doi: 10.1109/ickg55886.2022.00056

Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In *Proc. interspeech* (pp. 2843–2847).

Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2021). Converting foreign accent speech without a reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2367-

2380. doi: 10.1109/TASLP.2021.3060813

Zhao, G., Sonsaat, S., Silpachai, A., et al. (2018). L2-ARCTIC: A Non-native English Speech Corpus. In *Proc. interspeech* (pp. 2783–2787).

Appendices

A Metrics

The Word Error Rate is a metric used to measure the performance of automatic speech recognition systems. It is defined as follows:

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

where S is the number of words substituted by the ASR, D is the number of deletions (words omitted by the ASR), I is the number of insertions (words recognized by the ASR that were not in the original transcript), and N is the total number of words in the reference/label. The lower the WER, the better the performance of the model. WER is usually expressed as a percentage and can exceed 100% (the metric has a lower boundary of 0%).

The Character Error Rate is another metric used to evaluate the performance of automatic speech recognition systems. It measures the differences between the recognition results and the reference text at the character level. The formula for calculating CER is similar to WER but operates at the character level.

B Data Analysis

In this appendix, detailed tables illustrates the result of MOS from Experiment 1.

Table 3: Analysis of the MOS results from participants who are NOT familiar with Indian accents

ID	Comprehensiveness			Accentedness			Proficiency			Similarity	
–	SVBI	Conv	BDL	SVBI	Conv	BDL	SVBI	Conv	BDL	(1)	(2)
01	3.2	4.4	4.2	5	2.8	1.8	3.2	3	5	4.8	1
02	3.8	4.6	4.6	4.4	3	1.6	3.4	4.2	4.4	5	1.2
03	3.8	4.2	4.4	5	3.4	1	3.6	4.4	4.4	5	1.4
04	3.2	4.2	4.2	4.2	3.2	1	2.2	3.6	4.4	4.2	1
05	3.6	4.6	3.8	3.8	2.8	1	3	3.8	5	4.4	1
06	3.6	4.8	4.2	3.6	2.8	1	2.6	3	5	3.8	1.2
07	3.2	4.6	4.6	4.8	3.2	1.4	3.6	4.2	4.2	4.2	1.2
08	2.4	4.4	4.2	5	1.2	1.2	3.2	3.8	5	3.8	1.2
09	3.4	4.2	4.2	4.8	1.8	1	2.4	4	5	4	1.4
10	3.8	4.4	4.4	5	2.6	1	3.6	4	4.8	5	1.4
Average	3.4	4.44	4.68	4.56	2.68	1.2	3.08	3.8	4.72	4.42	1.2

Table 4: Analysis of the MOS results from participants who are familiar with Indian accents

ID	Comprehensiveness			Accentedness			Proficiency			Similarity	
–	SVBI	Conv	bdl	SVBI	Conv	bdl	SVBI	Conv	bdl	(1)	(2)
01	3.4	4.4	3.6	4.4	2.6	1	2.2	3	5	5	1.6
02	3.6	3.6	3.4	5	2.6	1.8	3.2	4.2	4.4	4.8	1.2
03	3.8	3.2	4.2	5	2.4	1.2	3.4	4.4	4.4	4.8	1.2
04	4.4	4.2	3.8	3.6	2.6	1.4	3.2	3.6	4.4	4.4	1.6
05	3.8	4.4	4	4.4	3.2	1.4	2.8	3.8	5	4	1.2
06	4.2	3.6	4.2	3.8	1.2	1.4	2.6	3	5	4.4	1.2
07	3.2	3.6	3.6	4.8	1.4	1.2	3.2	4.2	4.2	4.2	1.2
08	4.4	4.2	4.2	5	1.8	1.4	2.8	3.8	4	4.8	2
09	3.2	4.2	4	4.8	2.4	1.4	3.6	4	5	4	1.4
10	3.6	3.8	4.2	4.8	1.4	1	2.6	4	4.8	4	1.4
Average	3.76	3.92	3.92	4.56	2.16	1.32	2.96	4.22	4.36	4.44	1.4

Note:(1) pertains to inquiries involving two distinct audio samples from SVBI, where the task is to determine if they are spoken by the same individual. (2) involves presenting the original SVBI audio alongside its transformed version.

C Converted Speech Examples

The table below presents examples of Automatic Speech Recognition (ASR) results from the converted speech produced in Experiment 1. For comparison, the ASR results of the original CMU-ARCTIC(BDL) dataset are also included.

Table 5: ASR error examples of converted speech from Experiment 1

–	**Converted sample**	**ARCTIC-BDL sample**
arctic-b0440	THERE WERE STIR AND BUSTLE NEW FACES AND FRASH FRACTS	THERE WERE STIR AND BUSTLE NEW FACES AND FRESH FACTS
arctic-b0441	AND THERE WAS THE TAIL BAD WHOM ALSO YOU MUST REMEMBER	AND THERE WAS ETHEL BAIRD WHOM ALSO YOU MUST REMEMBER
arctic-b0442	HE HAD BECOME A MAN WEARY AND EARLY IN LIFE	HE HAD BECOME A MAN VERY EARLY IN LIFE
arctic-b0443	I DID NOT THINK HE WOULD BE SOLLY	I DID NOT THINK YOU WOULD BE SO EARLY
arctic-b0469	VERY FEW PEOPLE KNEW OF THE EXISTENCE OF THE SLOT	VERY FEW PEOPLE KNEW OF THE EXISTENCE OF THIS LAW
arctic-b0471	ALSO FEROSINADER CHANCE DE PE SAID SAID	ALSO A FELLOW SENATOR CHAUNCEY DEPEW SAID
arctic-b0475	NOT AVEALD MORE THAN HIS EMPIRE FIRE	NOT A WHEEL MOVED IN HIS EMPIRE
arctic-b0476	THE REORGANIZATION OF THESE COUNTRIES TOOK A FORM OF ROOLUTION	THE REORGANIZATION OF THESE COUNTRIES TOOK THE FORM OF REVOLUTION
arctic-b0477	YOU ARE GOING IN FOR GRAF SHERRY	YOURE GOING IN FOR GRAB SHARING
arctic-b0485	THE MOB CAME ON WHERE IT COULD NOUTER BONDS	THE MOB CAME ON BUT IT COULD NOT ADVANCE