



OILTHIGH
DHÙN ÈIDEANN

**ContinuousAccent:
An L2 Speech Dataset For
Foreign Accent Intensity**

B242890

7780 words

Master of Science
Speech and Language Processing
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
2024

Abstract

Foreign accented non-native speech is common in every culture around the world. The variability of non-native accent is poorly defined and inadequately measured. Because of poor understanding and labelling, it is near impossible to reasonably control non-native accent strength when synthesising speech. This paper builds a novel dataset of labelled foreign accent strength called *ContinuousAccent*. This research also uses several modelling techniques to determine that Whisper language identification is correlated with accent strength of speech. More work needs to be done to understand causation and to achieve generalisability of accent strength across language pairs. This research offers a dataset and baseline findings that can begin a path towards controllable foreign accent in speech synthesis and other related applications.

Acknowledgements

Thank you to my supervisor, Korin Richmond, my tutor, Siqi Sun, and all those that helped guide this research, Jinzumu Zhong, Noe Berger, and Adaeze Adigwe.

Also, thank you to my family who supported me, especially my wife, Victoria Hansen.

Table of Contents

1	Reflecting Natural Variation in L2 Speech Requires More Data	1
1.1	Motivations	2
1.2	Towards representation of accent diversity in synthesized speech . . .	3
2	Where Is Accent Intensity Modelling in the Literature?	5
2.1	Accent and L2 Speech	5
2.2	Current Work	7
2.2.1	Controllable Speech Synthesis	8
2.2.2	Automatic Speech Recognition (ASR) and Language Identifi- cation	9
2.2.3	Difficulty of Accent Modelling	10
2.3	Binary Classification Tasks	10
2.4	Phonetic Posteriorgrams	11
2.5	Whisper Language Detection	12
2.5.1	Conclusion	12
3	Using Models to Match What Humans Think	13
3.1	Data	13
3.1.1	SELL	13
3.1.2	Speech Ocean	14
3.1.3	L1 Data	14
3.1.4	Ideal Dataset	14
3.2	Data Collection for ContinuousAccent	15
3.2.1	Survey	15
3.3	Models	15
3.3.1	Mel Spectrogram Classification	16
3.3.2	PPG Analysis	16

3.3.3	Language Identification Model	17
3.3.4	Speech Rate	17
3.4	Hypothesis	17
3.5	Evaluation	18
4	Whisper Outperforms the Rest	19
4.1	Rater Agreement	19
4.2	Rankings	20
4.3	Correlations	22
4.3.1	Accuracy of Ranking	22
4.3.2	Accuracy of Discrete Labelling	23
5	Small Steps Toward L2 Accent Modelling	25
5.1	Limitations	26
5.2	Future Work	28
6	Research Summary	30
A	Additional Materials	37
A.1	Qualtrics	37
A.2	Qualtrics Results	39

Chapter 1

Reflecting Natural Variation in L2

Speech Requires More Data

Anyone who learns a foreign language understands that as you learn a language, there is an inherent difference between the way a learner speaks and the way a native speaks. When a native-born American learns French, French people may often describe the speech produced by this hypothetical language learner as an "Americanised" French. Foreign accented speech is very prevalent in today's speech corpora¹. Accents serve many functions. They can—for example—make a speaker stand out, as well as communicate ability or geographic information. The difference between native or first language (L1) and foreign or second language (L2) accent has produced a lot of literature in the field of language acquisition. There are differences between L1 and L2 speakers in terms of speech production as well as speech perception (Cutler, 2014; Porretta et al., 2016; Wieling et al., 2017). Foreign accents have long been identified not only by language learners and expert phonologists, but by nearly every human listener. This talent to hear accent is rooted deep in the human experience. In today's world of large data and artificial intelligence models, there is an appetite to mimic this human understanding and have automated metrics for evaluating accent.

Accented speech makes up a sizeable portion of our speech data. Williams et al. (2024) estimated that there is nearly three times as many L2 speakers of English as L1 speakers (1.08 billion versus 373 million). For example, in Common Voice, a popular dataset for training speech systems, it is known that a non-trivial amount of the data is L2 or non-native speech. Although L2 speech is prominent in our datasets, it's highly

¹Common Voice, Speech Accent Archive, and Arctic-L2 are all examples of datasets with foreign accented speech.

variable in accent realization and therefore difficult to identify, categorize, or even define. It is, therefore, unsurprising that researchers don't have a good understanding of (1) what defines an accent, (2) how to quantify the intensity of an accent, and (3) how to use this measure to inform practical accent-related tasks, particularly synthesising speech which accurately captures the complexity of natural foreign-accented speech. Given these limitations, the following work relies on the softly-defined linguistic concept of accent (See Section 2.1), and limits its focus on the problem of identifying the best automatic measure for accent intensity. We provide the following contributions:

1. A dataset, *ContinuousAccent*, containing human perception of accent strength that, to the best of our knowledge, is first of its kind.
2. An analysis of common automatic metrics and how they align with human perception of accent intensity. Each system is used to automatically evaluate accent intensity of unseen speakers and utterances on L1 and L2 speech. We explore the following methods for automatic evaluation:
 - (a) Binary classification using Mel spectrograms as input
 - (b) Binary classification using Phonetic Posteriorgrams as input
 - (c) Spoken Language identification task using a large pre-trained multilingual ASR model, Whisper
 - (d) and speech rate
3. A framework as to how these automated metrics could be used to inform speech synthesis systems to produce representative speech and accent controllable systems.

The main goal of this research is to model accent intensity. To be a good model of accent intensity, a system needs to be effective at matching human perception of accent strength. To achieve this goal, our study compares a range of speakers and ranks them. This research then uses models to evaluate how well each model ranking aligns with human perception.

1.1 Motivations

This research seeks to identify a metric that can help researchers quantify accent intensity and represent natural L2 speech. We hope that beyond representation of accent

intensity, this research can assist people in the domain of accent training. Accent training is when someone works with a professional language teacher or phonetician to improve their accent to sound less like an L2 speaker and more native-like. Having an automatic metric for accent strength would greatly improve the feedback loop and high costs associated with accent training and personal coaching.

L2 speech and foreign accent can come with negative stereotypes (Gluszek & Dovidio, 2010), so helping improve accent can certainly benefit individuals learning a foreign language. We believe that in speech synthesis it is beneficial to explore and represent natural variation. Greater representation of L2 speech in synthesised speech may help battle some of the stereotypes and stigmas associated with L2 accents or certain foreign cultures.

It has been shown that L2 learners perceive and understand L2 speech better than L1 speech (Munro et al., 2006). Thus, adjusting the accent strength of speech in commercial TTS systems may make the system more intelligible for L2 “non-standard” users.

1.2 Towards representation of accent diversity in synthesized speech

As we have experimented with speech synthesis systems, we have discovered that while research seems to be intensely focused on controllable text to speech (TTS), there are certainly limits on what is actually controllable. Modern systems are reasonably good at both identifying accent and controlling accent. For example, Melechovsky et al. (2022) and VALLE-X (Z. Zhang et al., 2023) have proposed a methods which allow for synthesising speech with English content in another arbitrary accent. However, there appears to be no regard for or control over the idiosyncrasies of natural variation in accent intensity. The current systems only understand the “stereotypical” accent. As with many features of speech, there is a natural distribution of the speech properties associated with accent. Think of someone you know who speaks your language as a second language, then think of their mother or grandmother, whose language is hypothetically much more accented. Research still has a lot of ground to cover to understand and control this variability in accent. The current distribution of synthetic L2 speech is imbalanced and doesn’t represent the uniqueness of the real world distribution. This shows that current models are far from modelling human perception and

production of natural speech. Understanding this variability in L2 accent could assist systems to improve language understanding and recognition as well as inform speech synthesis tasks.

Chapter 2

Where Is Accent Intensity Modelling in the Literature?

2.1 Accent and L2 Speech

Speech that is produced by second language learners (L2 speech) differs greatly between native and non-native speakers. L2 speech is “loaded” with phonological and other information about the first language of the speaker (Cutler, 2014; Ordin & Polyanskaya, 2015). This phenomenon is often referred to as *interference* in the second language acquisition literature (Bhela, 1999; Derakhshan & Karimi, 2015). Interference patterns are unique to the speaker’s L1, as evidenced by researchers dedicating their time to the problem of *native language identification* (NLI). NLI is a automated task which identifies a speaker’s native language by listening to their speech in a second language. Thanks to modern neural networks and compute resources, there have been several successful models that can perform NLI (Graham, 2021; Humayun et al., 2022). This evidences the existence of accent or the patterned difference between L1 and L2 speech.

The term *accent* often refers to the overall variation in speech production and pronunciation (Acheme & Cionea, 2022; Markl & Lai, 2023). Accent is always defined from the perspective of a single listener and how that listener perceives differences from their own speech. Because of its extremely subjective nature, accent is easy to define generally but hard to define specifically. For example, two native speakers of General American English might perceive different accent intensities from the same German speaker depending on their individual exposure and experience with Germans or the German language. It is known that an increase in language input is a driving

factor in L2 speaking proficiency and L2 language perception, and even L1 perception (Gass et al., 1998; Saito & Hanzawa, 2018). L2 accented speech, depending on the listener, can differ in pronunciation, and potentially even things like class or status (Markl & Lai, 2023). This subjectivity makes the specifics of accent definition extremely hard. More formally defined, an accent may be identified as “foreign” if the speech features diverge from that of the listener’s systematically at the phonetic, phonotactic, phonological, lexical, segmental, or suprasegmental levels (Cristia et al., 2012; Wells, 1982). So, although accent is generally regarded as a phonetic phenomena, it is more than that. Accented speech is often classified by a listener as a “standard” or “non-standard” variety (Acheme & Cionea, 2022; Markl & Lai, 2023). There is a defined standard or stereotypical Swedish accent, for example. Standard accents adhere to codified norms which are familiar to the listener and can be easily grouped together by similar properties. Non-standard accents depart from these codified norms and may account for idiosyncratic variations. (Dragojevic et al., 2018).

Accent strength of speech is also somewhat subjective, however, it seems to be generally understood as the degree to which the speech differs from “standard” or “native” speech. There seems to be a strong consensus that accent is something that is perceived in discrete degrees of strength (Piske et al., 2001).

Many researchers have investigated the reasons for the variability seen in the degree of accent. These reasons include demographics, history, age of acquisition, L2 exposure, and perception of one’s identity, just to name several (Marx, 2002; Piske et al., 2001). Many studies have also tried to quantify this degree of variability most often using a rating system—like the Likert scale—which compares speakers—usually learning English—on the sentence level (Jesney, 2004). There is good reason to critique the Likert scale for rating accent, as it is not explainable; there is no clear way to define what a 1 accent is compared to a 5 accent. In this case, individual perception and gradients are also not comparable. Despite this, the use of the Likert scale in global foreign accent rating is popular and, by many, deemed a reasonable approach (Jesney, 2004). Because of its difficulty to define, it is common to discretise accent strength and define categories (i.e. scales from 1-5 or strong to weak), however, some research has been done to quantify accent and accent strength as a continuous value (R. Liu et al., 2024; Zuluaga-Gomez et al., 2023). The purpose of *ContinuousAccent* is focused on making accent intensity both a continuous and learnable representation extracted from speech signals.

In the recent era of large models and big data, accent intensity is an appealing prob-

lem because its definition is not-concrete yet accounts for a large variation in speech. Can we learn a good representation of L2 accent with a large model and lots of L2 speech data?

2.2 Current Work

This study will look at current methods which can proxy accent intensity. There have been several recent methodologies for trying to control accented speech, mostly for speech synthesis purposes. One methodology is to train a neural network—or learn—a representation of speaker identity. One can then condition speech synthesis on that speaker ID so that the output speech sounds like that person (Chen et al., 2019; Ma et al., 2023; Shimizu et al., 2024), however, this is problematic as accent and speaker ID are clearly entangled. Taking it a step further, another method is to disentangle accent and speaker identity, and then at synthesis time prompt with speaker identity and accent ID (Ding et al., 2022; Peri et al., 2020). However, no one has tried to synthesise variable L2 speech. The focus seems to be on *Foreign Accent Conversion* and stripping the L2 features to produce L1-sounding speech while maintaining the speaker’s identity. Experiments have been run on mixing accent ids to get accent gradients (Y. Zhang et al., 2019). Lecumberri et al.’s (2014) notable work used HMM-based models and blending voice models for the L1 and L2 language to generate degrees of accented speech. This study proposed several generative approaches to alter speech, which they claim resulted in “a reasonably convincing degree of foreign accent for consonants.” One problem with this method is there is little justification—besides “it works”—for mixing the models. In other words, there is no explicit modelling or learned system that understands what of foreign accent is and what parts of the data control its variation.

There is some recent work being done that is trying to directly model accent intensity (R. Liu et al., 2024). In this paper, there was a distinction between phoneme level accent features and utterance level features of accent. The model, we refute, doesn’t have a true definition—or ground truth—of accent to properly learn accent intensity. Thus, the need for a better dataset. They proxy accent intensity with a Goodness of Pronunciation (GOP) model (more about GOP in 2.2.2). Additionally, R. Liu et al.’s (2024) evaluations seemed to be inconsistent (listeners with different language backgrounds). Furthermore, the evaluations seemed to be based on the perception of accent of synthesized speech (conditioned on their unbased 0.1-0.9 accent rating). We argue that this approach is basically changing the f0 frequency and not the underlying accent

intensity. The generated speech in Lui’s work is far from expectations for accent control. We refute the claim that this model has “good expression” of intensity (R. Liu et al., 2024).

If you want to synthesise speech, you need to be able to have data that is labelled along the dimension you are wanting to control. Because of this, we hope that *ContinuousAccent* and similar datasets will enable controllable speech synthesis.

2.2.1 Controllable Speech Synthesis

Currently, there are several types of speech synthesis systems that aim to control accent:

- Foreign accent conversion (FAC) is a popular task in speech synthesis. FAC is a model which inputs accented speech and outputs speech with “no” accent, or another predefined accent. There have been several models that learn accent representations, and condition synthesis on accent ID or learned embedding to achieve voice conversation like Accentron (Ding et al., 2022) and others like (S. Liu et al., 2020). These methods change the conditioned accent, but do not understand anything about features or strength of that accent.
- A few researchers have developed speech synthesizers that are conditioned on phonemic information or Phonetic Posteriorgrams (PPG) or pronunciation information like (Churchwell et al., 2024; Zhao et al., 2019). By changing the PPG representation, they can control pronunciation during synthesis, but not overall accent.
- Another approach researchers have taken is to condition synthesis on speaker identity. This is a form of voice conversation or mimic. This task can be used to mimic the accentedness of one speaker with the voice of a separate speaker. This has been experimented with simple KNN acoustic feature selection as in (Baas et al., 2023, Preprint), as well as end-to-end systems like Parrotron (Biadsky et al., 2019) or older HMM systems like (Sisman et al., 2020; Ye & Young, 2006). It is important to understand here that speaker identity is very much entangled with accent. In order to model accent/accent intensity, it should be disentangled from the speaker identity during model training. This is a difficult task since no datasets exist of a single speaker speaking in many different accents or accent strengths.

2.2.2 Automatic Speech Recognition (ASR) and Language Identification

ASR benchmarking is often used as a proxy for intelligibility. However, We think it is important to understand the unique challenges these metrics present when working with accented L2 speech (Aksënova et al., 2021). Foreign accented speech is difficult because the variability of accent is large, “complex, and spans multiple dimensions” (Aksënova et al., 2022, Preprint). For example, foreign accent can vary in phoneme patterns, but also common grammar and vocabulary mistakes. This provides a challenge to ASR systems that rely heavily on language modelling and assume correct grammar and vocabulary. Thus, the language models will optimise for what the speech meant to say instead of what was actually said.

Much of the research in TTS uses ASR WER as an automatic and objective metric in order to measure intelligibility of speech (Cumbal et al., 2024; R. Liu et al., 2024). These large ASR systems often are worse than humans at recognising speech, because humans are far better at dealing with accents, noisy environments, and varied speaking styles (Scharenborg, 2007). However, automatic systems are more more efficient than humans in identifying languages which would otherwise be unfamiliar to the listener.

Another common downstream task of ASR systems is language identification. These models input speech and then output a probability distribution of languages. This type of task has been used to proxy accent strength of a speech signal (Kukk & Alumäe, 2022; Tännander et al., 2024). This proxy assumes that heavily accented speech is more likely pick up features of the L1 and distribute that probability mass to the L1. This method for getting a posterior probability of languages has been used to make claims about accent strength in English-accented Swedish speech (Tännander et al., 2024).

ASR systems are also fundamental in creating systems in computer-assisted language learning, specifically delivering feedback about pronunciation error. These systems are referred to as Goodness of Pronunciation (GOP) systems (Kanters et al., 2009). Although accent and pronunciation are two different language phenomena, pronunciation is a factor that plays a part in accent intensity, thus GOP systems could potentially prove useful in accent intensity modelling. However, R. Liu et al. (2024) attempted accent intensity modelling with a GOP system which didn’t provide control over the accent.

Other non-ASR models have specific training for NLI such as Ubale et al.’s (2018)

Listen, Attend, Identify or a preceding model from Qian et al. (2017). These models could potentially prove insightful and/or be adapted to the task of rating accent intensity, similar to the language identification task that ASR systems employ.

2.2.3 Difficulty of Accent Modelling

As explained earlier, one of the main difficulties associated with accent intensity modelling is that we don't have concrete definitions for accent and the specifics vary depending on language and the identity of the listener. Accent is too subjective. Its subjectivity makes it difficult to collect data that is both consistent across speakers and controlled from a listener's perspective, which is difficult even from people of similar L1 language backgrounds. Because of this loose definition, current systems proxy accent with pronunciation or intelligibility.

Another major difficulty in accent modelling is there is low data availability. There is a lack of data which represents a variety of accent strength from a controlled speaker. Having this variety would require recordings of a person speaking a foreign language from a heavy accent until they are near native. Not only would this dataset be expensive and difficult to collect, but it would likely span over years with no guarantee of accent intensity improvement. There is also a general lack of datasets containing foreign accented speech. To the best of our knowledge, there are no datasets that have labelled accent strength.

2.3 Binary Classification Tasks

Accent modelling, we believe, justifies a reasonable use case of binary classification. A speech signal can be L2 accented or not accented. A binary classification model is one whose output is a number between 0 and 1, each respectively representing one of these classes. The actual model or function that turns the input into a $[0, 1]$ representation is arbitrary to the task. It can be any type of neural network or just simple mathematical operations. The last transformation in these types of models—often called the activation layer—are functions that take any number and force it between 0 and 1. Two popular activation functions in neural networks with this constraint are sigmoid and softmax functions. Sigmoids are often used for discrete labels 0 and 1, whereas the softmax function is often interpreted as a probability distribution. Thus, in the case of accent, where accent strength in speech is certainly not a discrete value, a softmax

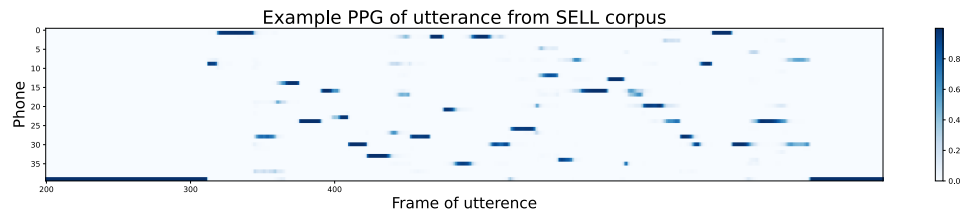


Figure 2.1: This figure is a visualisation of an entire utterance “Dad’s going to be somewhat peeved if we lose that boat” from the SELL corpus. PPGs detail the frame-wise distribution of phone predictions. The cross sections with dark blue show more certain predictions, and the sparse areas with lighter blue indicate less certain predictions of the phone represented in that part of the speech signal.

activation is a more appropriate method.

2.4 Phonetic Posteriorgrams

A phonetic posteriorgram (PPG) is a frame-by-frame posterior prediction of a phone class. These phone classes come from a set of predefined phonetic units, which generally retain the linguistic and phonetic information of the utterance. So, the basic goal of a PPG is to map each frame of speech to a probability distribution over the phoneme categories. This gives a map which shows the most probable phonemes across the speech signal. The reason that this is interesting for our work on accent intensity identification is because prior work has shown that PPGs of L1 and L2 speech are significantly different (Youngsun & Hosung, 2021). This difference suggests that PPGs could encode information about L1-ness of an utterance. Under the assumption that the variability of L2 accented speech is encoded within a PPG, then this method could potentially be used to synthesize variable accented speech. To our knowledge, prior work on PPGs has not been used on accent intensity tasks or accented speech. However, prior works have shown that PPGs are able to synthesise speech and convert between foreign accents. This is possible by training a model to convert PPGs to Mel spectrograms and then using a standard vocoder to output a playable waveform (Zhao et al., 2019).

2.5 Whisper Language Detection

The development of large pre-trained models, such as Whisper, has recently been popular. There have been many researchers that have used these systems to also perform a task often referred to as language identification or detection (D. Liu et al., 2021). This is a natural problem for ASR systems, because in order to recognise speech, the first step is to understand the language the speech is spoken in to take advantage of knowledge from a known vocabulary and grammar. One of the largest and most widely used ASR systems is called Whisper (Radford et al., 2023). Whisper is from a company called OpenAI, and is an open-source—and most importantly—a multi-lingual model which has language detection capabilities of 100+ languages.

In Tånnander et al. (2024), the authors report an objective measure of strength of English-accented Swedish using `whisper-large-v3`. This large model is used by the authors to calculate the probability of a language. Afterwards, they report the probability of English as the strength of the accent present in Swedish utterances. These results indicate that this objective measurement could perhaps correlate with accent strength. However, this is not a tested assumption as accent strength has not been directly measured using `whisper-large-v3`.

2.5.1 Conclusion

In summary, accent modeling remains a fascinating, challenging, and under-researched task. Foreign accented speech is both deeply subjective, yet codified and patterned. While substantial progress has been made in areas such as accent conversion and speech synthesis, significant obstacles persist, particularly in defining and quantifying accent strength. Current methodologies often rely on proxies like pronunciation or intelligibility, additionally there is a general lack of concrete definitions and controlled datasets in accent modelling. Despite these challenges, classification models, PPGs, and large pre-trained models—like Whisper—offer potential directions for future research. By improving accent modelling, we can ultimately contribute to advancements in accent training, language learning, and speech technology.

Chapter 3

Using Models to Match What Humans Think

The goal of our experimentation is to understand human perception of accent intensity through data collection and analysis. We surveyed several current tools to understand what methods capture accent variability enough to model accent intensity. In this section, we will discuss the data both used and collected for experimentation, as well as the tools used to proxy accent intensity.

3.1 Data

In order to get human perception of accent, we had to find a corpus of foreign accented speech that had a few constraints: (1) the L1 and L2 needed to be the same for all speakers for simplicity, (2) there needed to be sufficient variation within the accent intensity across the data set, and (3) sufficient data was needed to train simple models.

Firstly, there are not a lot of speech datasets that have foreign accented speech data from people of the same L1 and L2 background. That greatly narrowed the available datasets. The two largest datasets that met constraint (1) were the SELL Corpus (Yu Chen & Zhang, 2019) and the speechocean762 corpus (J. Zhang et al., 2021).

3.1.1 SELL

The SELL corpus is a multiple accented speech corpus for several regional dialects in China. All the speakers are L1 Chinese and L2 English speakers. It is open source and has 31 hours of speech. This was an interesting corpus, however, through listening

through many of the speech, there was very little perceived variation in the accent. Yu Chen and Zhang (2019) acknowledge that “further studies and data collection are necessary to bring a deeper understanding on accents and pronunciation errors by L2 speakers”. This disqualified the corpus from use for human perception of variation, however, the Mandarin Chinese data is useful for training models on L2 speech.

3.1.2 Speech Ocean

Speechocean762 (speechocean) is a dataset that is comprised of 250 different speakers from various backgrounds. All the speakers, however, are L1 Mandarin Chinese and L2 English speakers. This dataset’s variation in accent was more obvious, making it a more ideal dataset to use to understand and test human perception.

3.1.3 L1 Data

In order to train our classification models, we needed L1 speech to balance the classes for training. We tried creating a dataset composed of parallel utterances from the L2 corpora (SELL and speechocean) using ParlerTTS (Lyth & King, 2024), but realized that having the same sentences was not necessary, so we decided to use more authentic L1 data. We used TIMIT, because the data comes completely from 7 American dialects of English. Only 5 accents were used to balance the dataset between L1 and L2. More importantly, the TIMIT data is also short utterances of English sentences. This matches the data in speechocean. That is especially important for the PPG task, we didn’t want the number of frames in the L1 and L2 PPGs to differ dramatically.

3.1.4 Ideal Dataset

The perfect dataset would include a controlled group of people of the same L1 speaking in their L2 language(s) over a long span of time to improve their accent until they sound native-like. Unfortunately, a dataset of this nature would be very expensive and difficult to procure.

3.2 Data Collection for ContinuousAccent

3.2.1 Survey

In order to collect human perception of accent strength, we used the Qualtrics survey platform to ask native English speakers to rank speaker utterances in terms of accent strength. There were 2 rounds of surveying. In the preliminary round, participants were selected from the network of the researcher. In each question, there were 3 utterances from random speakers of our dataset (See Appendix A.1 for more information about the Qualtrics survey). We constructed 500 random comparisons of speakers that covered the entire dataset. We fixed 2 random questions, which every participant saw, and randomly selected 23 more from the now 498 question bank. The preliminary test had 25 questions. These 2 fixed questions enabled us to test our confidence in inter-rater agreement while focusing on collecting as many comparisons as possible. After collecting preferences, we ranked the speakers in terms of the accent intensity (see section 3.5 for discussion of the ranking method). After this initial ranking, we used the results to construct a more discriminative survey to send out using Prolific—a survey recruiting platform. This discrimination was done such that each speaker in the final survey was not compared against a random person in the dataset, but against those that initially ranked between 5-10 places higher and lower. A 5-10 ranking difference was chosen because it would allow for the task to be difficult enough to be meaningful but easy enough for participants to do quickly. Ranking two speakers that are the same accent strength is more difficult, but more helpful in increasing ranking confidence. So we wanted pairs that initially ranked close to each other, but not too close. Ethical approval was granted by the an ethics committee before results were collected. These additional results added to the preliminary results and gave us a final human perception ranking of speakers from best to worst accent.

3.3 Models

For the experimentation, we built 2 models: one convolutional neural network (CNN) classifier conditioned on Mel spectrograms and one feed-forward neural network (FFNN) classifier conditioned on a PPG feature vector. Because we were working with a relatively small dataset, we needed to keep the models small as well to avoid overfitting, for this reason, our 2 models were only 2 simple neural network layers plus an activation

function.

3.3.1 Mel Spectrogram Classification

We implemented a simple binary classification model that outputted a probability of L1 or L2 speech. The input to this simple model was a regularized Mel spectrogram obtained from the popular python package `librosa`. We used a small CNN that models a regression task to predict the probability of L1 speech given the Mel spectrogram. We used the dataset of L1 and L2 speech TIMIT and SELL corpora. We normalised each spectrogram to a fixed size before feeding it into the network. We then used a softmax function to get the probability that the signal is L1 vs L2 speech. Because there is not a strict decision boundary between L1 and L2 speech, we increased the temperature ($\tau = 2$) of the classification task to have a higher probability of predicting intermediate values between 0 and 1.

Other models like Lesnichaia et al.’s (2022) model to classify foreign accent, use 2D convolutions on spectrograms. A 2D CNN treats Mel spectrograms as images, which implies that the signal has invariance with both time and frequency. Because of the simplicity and commonness of Mel spectrogram CNNs, we decided that it would be a good baseline model of accent strength.

Using our speechocean dataset, we produced normalised Mel spectrograms and then ran inference on our trained model. We ranked the speakers by average output probability of L1 across each speaker’s utterances.

3.3.2 PPG Analysis

Because accent is generally observed to be a phonetic phenomena (See Section 2.1), we wanted to use models that had lots of phonetic context to improve our prediction accuracy. We used PPG representations to model the differences between L1 and L2 speech. Firstly, we needed to standardize the input. Because PPGs are frame-wise matrices and different utterances have a different amount of frames, there is an inherent alignment issue. We decided to sum over all the frames of the PPG. This summation essentially provides a feature vector, which can be used as a standard input to any regression function. We made a conscious decision to sum frames rather than average over the frames, because we wanted to keep effects of duration.

After normalising for length of PPG as input, we used 2 linear layers with ReLU activations in a FFNN to produce a binary classification task. This network is trained

using the same L2 data from SELL and L1 data from TIMIT to predict each PPG as L1 or L2.

The PPGs were generated by an open-source PPG model and python package called `ppgs` that was developed by Churchwell et al. (2024, Preprint). The model uses the standard 40 phone set of phonemes as defined by Arctic CMU (Kominek & Black, 2004).

Using our speechocean dataset, we formatted PPG feature vectors and then ran inference on our trained model. We ranked the speakers by average output probability of L1 across each speaker’s utterances.

3.3.3 Language Identification Model

Whisper is a very popular open-source software. We used the `large-v3` model of Whisper (see section 2.5). We used the python package `whisper`¹ to run our speechocean corpus through the `model.detect_language` function and extract probabilities of English and Chinese. We used these probabilities to rank the speakers from highest probability of English to the lowest averaged across all their utterances.

3.3.4 Speech Rate

We also wanted to test the ability of speech rate to predict accent strength. Munro and Derwing (1998) found that “listeners evaluated passages read slowly as more accented”. If speech rate is an indication of accent strength, that is a simple calculation that doesn’t have to be learned. We calculated a speaker’s speech rate by taking the length of audio and dividing by the number of words in the utterance, averaged over all utterances. Speakers were easily sorted by speaking rate to model most to least accented.

3.4 Hypothesis

We expected that the phonetic information captured by ASR tasks will be more informative than Mel spectrogram for accent intensity modelling. So, we expected Whisper and PPGs to outperform the Mel spectrogram classification and speech rate. We hope that the learned projection of the PPG modelling will be most informative since it will have learned specifically from our accented dataset. We note that we don’t actually

¹<https://pypi.org/project/whisper/>

know what data Whisper was trained with, but it is reasonable to assume that some accented speech is present in the 5 million hours of training data ².

3.5 Evaluation

The goal of these experiments was to align the output of the models to fit the human perception. We constructed a rank of human perception, and we collected similar ranking by running utterances from the speaker through the models to obtain a ranking. The overlap of these rankings will inform us of model ability to match human perception of accent strength.

The rankings for human evaluation will be determined by TrueSkill ranking (Herbrich et al., 2007). TrueSkill was developed by Microsoft for ranking players in video game leader boards. It uses a Bayesian inference algorithm to calculate likelihood of ranking based on performance and it updates these statistics after each observation—in our case, pairwise comparison (is speaker A better than speaker B?)—to reflect the new information. Unlike traditional ranking systems, TrueSkill will account for uncertainty in a speakers ability, making it an attractive ranking algorithm for this use case.

An important aspect of these experiments is to understand the similarities between two sets of ranking. A simple Pearson rank correlation is used to compare our human evaluations and model outputs. Pearson correlation is a great way to understand the similarity between 2 numerical lists. Because of this numerical constraint, we converted every ranked list of speakers into lists of IDs from 1 to 250.

Because accent strength is typically discretised, we also measured accuracy of placing each speaker in the correct group. If the lists are categorical, we can measure (1) the accuracy of assigning the right label, and (2) a composite score of precision and recall called the macro F1 score. We used macro F1, as opposed to micro or weighted F1 in order compensate for the class imbalance. We experimented with assigning 2 categories (strong and weak accent) and 3 categories (strong, weak, and neutral/slight accent).

²<https://huggingface.co/openai/whisper-large-v3/blob/main/README.md>

Chapter 4

Whisper Outperforms the Rest

We received 40 participant responses from the preliminary survey which was enough responses to have full coverage of our 250 speaker dataset. Using Prolific, we collected responses from 5 groups of 5 American English speakers to answer 50 curated questions each. This again ensured full coverage of the 250 speaker speechocean dataset.

4.1 Rater Agreement

Firstly, the control questions (See Section 3.2.1) in the preliminary survey indicated that there is general agreement between raters. We used Kendall’s tau (Knight, 1966) to calculate the inter-rater agreement because of its ability to process larger sets of rankings.

Question	Agreement
Q16	0.7586
Q330	0.5361

Table 4.1: The two control tests show there is strong and moderate agreement between the raters. This shows that we can likely trust the rankings of the raters. Note: 0 is no agreement, and 1 is perfect agreement. Question 16 and 330 were randomly selected from the bank of 500 questions in the preliminary survey. See Appendix A.1 for more information about the survey results.

4.2 Rankings

As a result of the human evaluation survey, we had 1860 pairwise comparisons. Using the TrueSkill algorithm (See Section 3.5) we probabilistically ranked each speaker and visualised the distribution of accentedness (See Figure 4.2). This ranking algorithm provided a score and, more importantly, a ranking of all the 250 speakers. After the preliminary survey, we observed that the accent scores from TrueSkill pass a Shapiro-Wilk normality test ($p = 0.787$)¹ providing some evidence that accent strength is a normal distribution (Shapiro & Wilk, 1965). After collecting all the human perception data, the Shapiro-Wilk test continues to show normality ($p = 0.251$). Figure 4.1 shows that the output probabilities of our models (Mel spectrogram classification, PPG classification, Whisper language identification, and average speech rate) have non-normal posterior distributions. This mismatch of score distribution shape may suggest that these models are not appropriate for accurately scoring accent intensity, but can help us with ranking speakers.

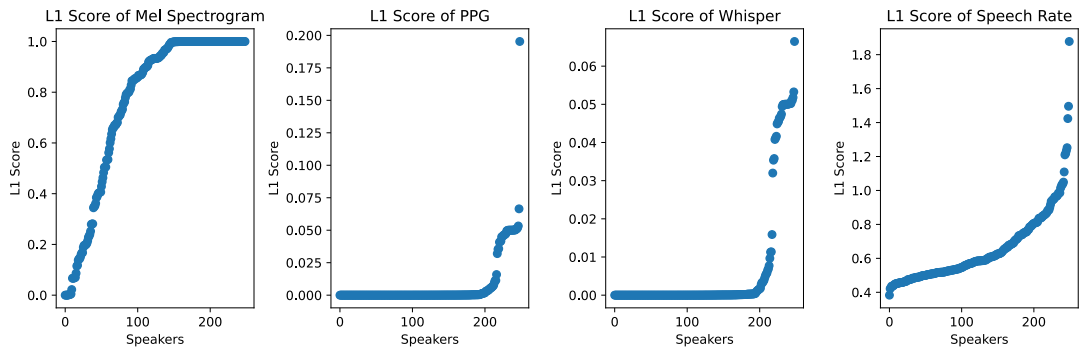


Figure 4.1: The score rankings of the Mel spectrogram, PPGS, Whisper, and speech rate models are significantly different from a normal distribution (Shapiro-Wilk p-values $3.15e^{-19}$, $5.47e^{-28}$, $4.36e^{-27}$, and $4.86e^{-16}$ respectively).

¹The null hypothesis of the Shipiro-Wilk normality test is that the distribution is normally distributed. So, a p value below the threshold indicates non-normality.

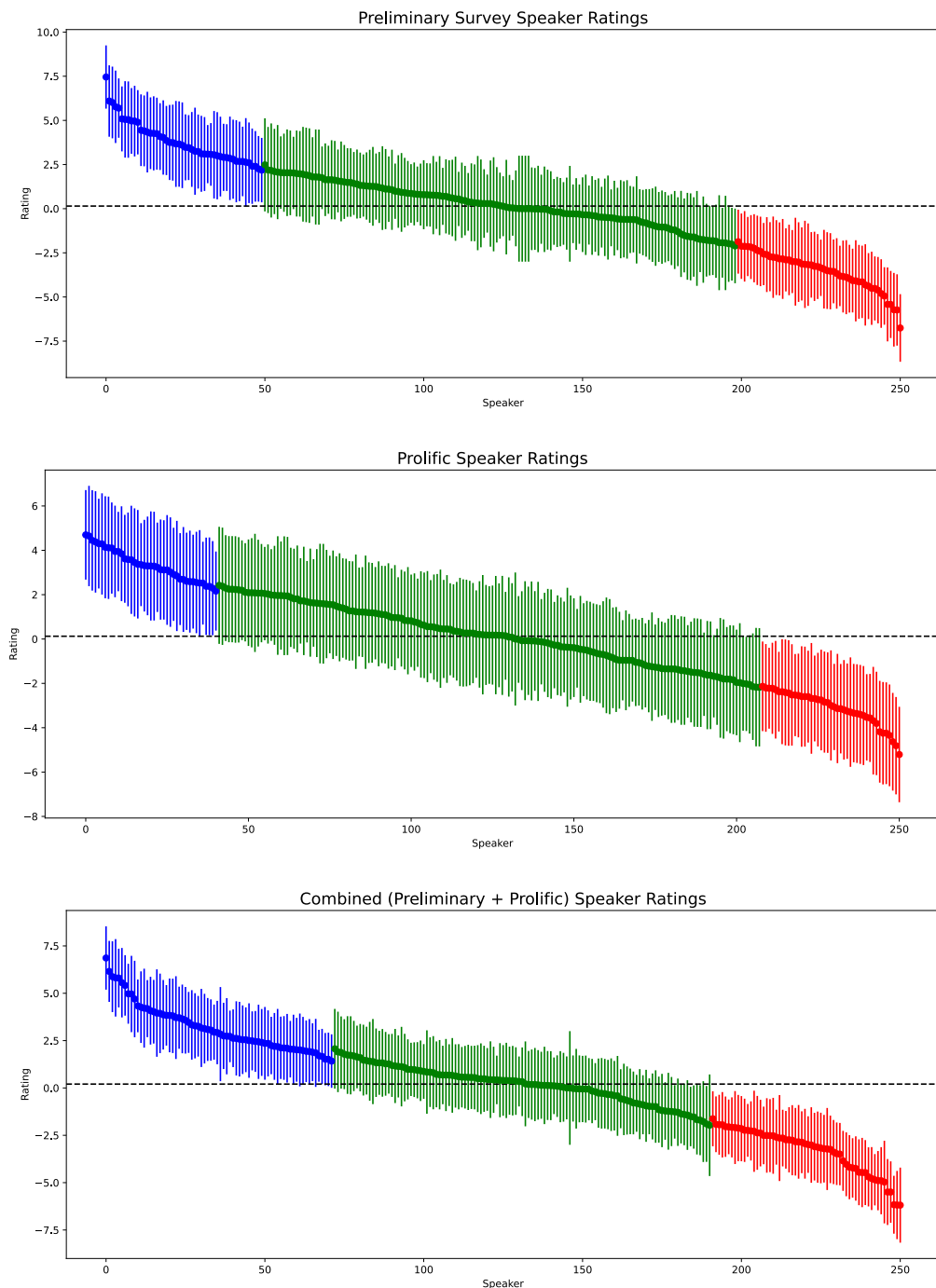


Figure 4.2: TrueSkill scores with error bars on preliminary survey, prolific survey, and the combined data set respectively. High rating indicates weak accent, and low rating, strong. In the case of 3 discrete labels, weak accent was defined as significantly above the average and strong accent was defined as significantly below the average. For the case of 2 labels, weak accents scored above the average and strong accents scored below the average. Average score is marked with a black dashed line.

4.3 Correlations

4.3.1 Accuracy of Ranking

We determined the correlation between each of the ranking algorithms: human perception, Mel spectrogram classification, PPG classification, Whisper language identification, and speech rate. The highest correlation with the human rankings is OpenAI’s Whisper (See Table 4.2 and Figure 4.3). The second highest correlation is between human perception and speech rate. We ran the correlation statistics for all models against each other, interestingly, Whisper and speech rate are the next highest correlation even though it is weak correlation.

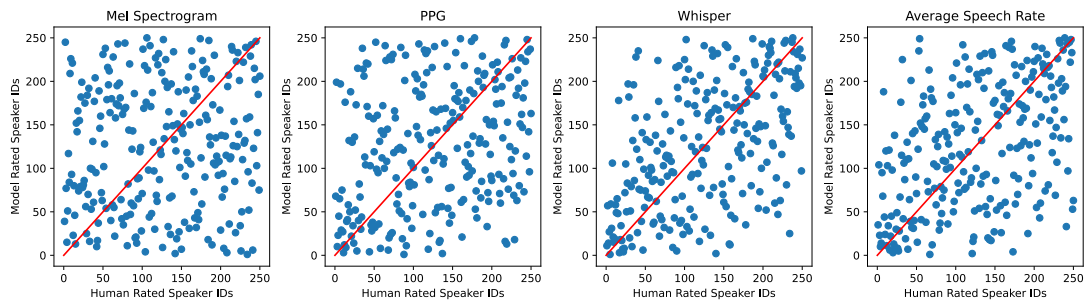


Figure 4.3: Visualisation of the Pearson correlation metric for Mel spectrogram, PPG, Whisper, and speech rate conditioned rankings. A perfectly aligned list is shown as the red diagonal.

	Binary	PPG	Whisper	Speech Rate
Human	0.0849 ($p = 0.1808$)	0.2697 ($p = 1.5357e^{-5}$)	0.5168 ($p = 1.7657e^{-18}$)	0.4473 ($p = 1.055e^{-13}$)
Binary		0.0712 ($p = 0.2614$)	0.0056 ($p = 0.9289$)	0.0893 ($p = 0.159$)
PPG			0.2101 ($p = 0.0008$)	0.0757 ($p = 0.2328$)
Whisper				0.2709 ($p = 1.401e^{-5}$)

Table 4.2: Correlation matrix for all methods of ranking speakers on accent strength. Pearson correlations given with their respective p-values. Whisper is best correlated with human perception, followed by speech rate, and then our PPG model.

Model	Accuracy	F1 Score	Accuracy	F1 Score
Binary	0.512	0.49782	0.396	0.36393
PPGs	0.592	0.58015	0.428	0.40726
Whisper	0.72	0.71186	0.592	0.59116
Speech Rate	0.704	0.69540	0.56	0.55778

Table 4.3: Grouping of 2 strong, weak (left) and grouping of 3 strong, neutral, weak (right). Accuracies were generally higher in groupings of 2, but less meaningful in context.

4.3.2 Accuracy of Discrete Labelling

The Pearson correlation allows us to see the accuracy of the exact ranking of the speakers using each of the modelling tasks. This is nice, however, exact order may not be important in the context of accent, thus the popularity of discretisation. We also want to see how accurately these models assign the right label (ex: speaker1 = 'strong' accent). To do this, we have all rankings from all models, so we assign the discrete labels based on the posterior distribution of the human rankings list. In other words, the distribution of categories for each model ranking is the same as the combined ratings from Figure 4.2. This shift from continuous to categorical, helps improve the accuracy of the system. We observed that Whisper was still the highest performing model. Surprisingly, it was closely followed by speech rate (See Table 4.3). As noted in Section 3.5, the classes are imbalanced, so in order to visualise the accuracy of these models, we plotted confusion matrices and plot the precision of each true label class in Figure 4.4.

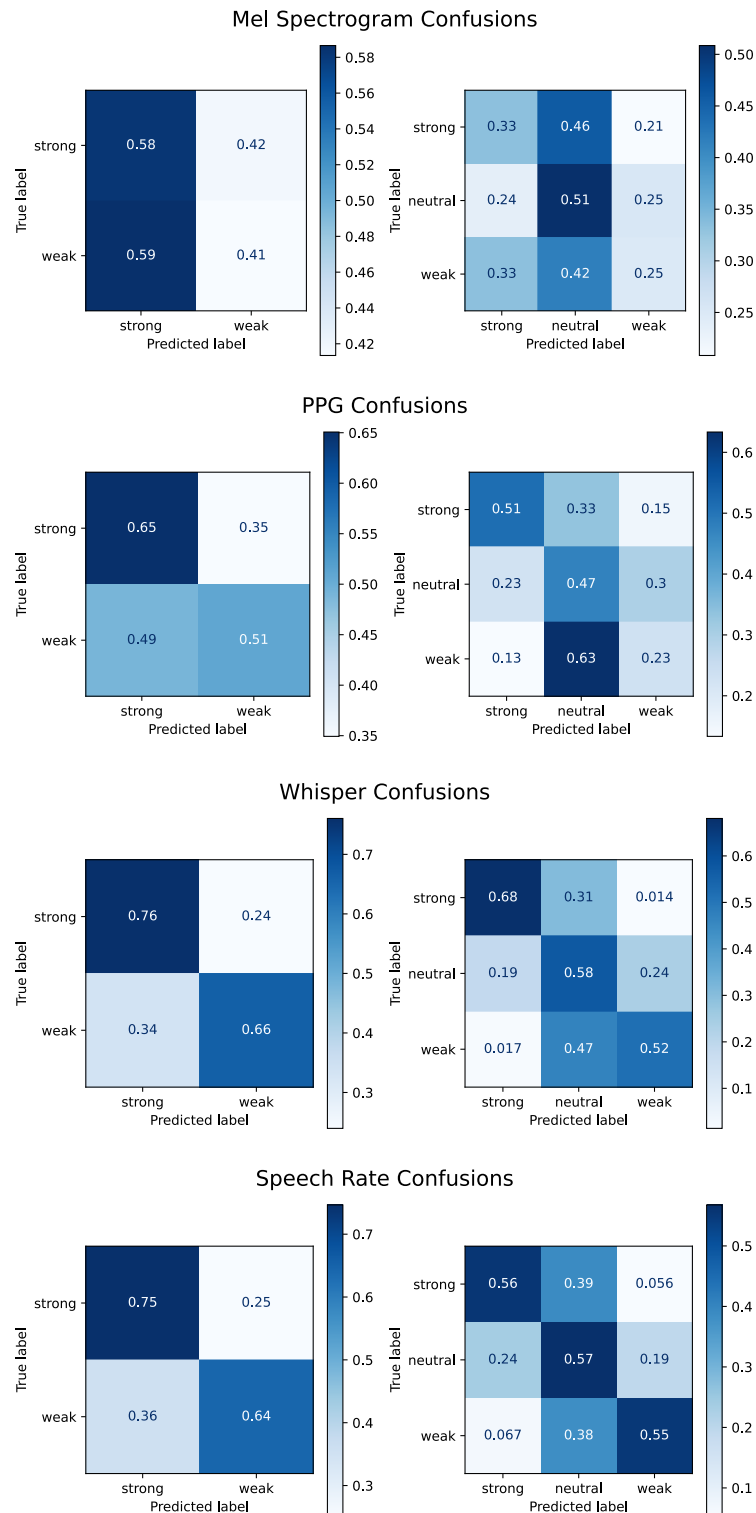


Figure 4.4: Confusion matrices for each model's accuracy in determining accent strength as a discretised label. Whisper and speech rate seem to be the best at discriminating strong from weak and weak from strong as indicated by the stronger diagonal and low scores on the negative cases.

Chapter 5

Small Steps Toward L2 Accent Modelling

This problem of trying to label accent strength is a perfect illustration of the one-to-many problem of speech synthesis (Godoy et al., 2009; Mohammadi, 2015). Any given text can be spoken as L1 speech, or L2 speech from any L1 background with any degree of accent strength. This number of possible utterances from one text input is already extremely high if you consider all the languages in the world and just a limited number of accent strength categories. This number is many times greater if you consider all the prosodic variations of speech. All these variants could come from just one text input. This problem is what makes speech synthesis a difficult and interesting problem. Labelling speech with L1 and L2 and accent strength is a way to add features that change the problem from one-to-many to one-to-fewer. By labelling the speaker utterances of the speechocean dataset with information about the human rated accent evaluation, we can better inform synthesis that happens in regards to the variation of accent. This dataset is called *ContinuousAccent*.

In this study, we found that the language identification module of Whisper is the highest correlated metric to human perception of accent strength. The next highest correlation is speech rate. Because the posterior distributions of Whisper scores and speech rates are different and their correlations are weak, it is possible that a combination of these models will provide a performance gain. It is often the case that a combination of models can generally outperform any one model (Weigel et al., 2008).

In many ways, the discretization of accent strength makes sense, because the difference between the best and second best accent strength in a large corpus is negligible compared to the best and the worst. Thus, we don't want our results to be punished

based on small differences, but large differences. So, if discretization makes sense, then the main concern becomes, how many accent strength classes are appropriate? 2? 3? 10? This is the main argument for continuous variables. We want a continuous representation of accent strength rather than an ambiguously defined 'strong' or 'weak' accent.

In this paper, we propose a method for understanding the efficacy of a given model of accent intensity. Our findings are that Whisper has the best correlation with human rankings, however, this is not causation. We cannot say that Whisper language identification is a good model of accent intensity, only that the two correlate. There are many reasons why this could be. There are 1.55 billion parameters in `whisper-large-v3`, which could account for this increase in accent strength discrimination power. It is hard to be certain without a good understanding of the data used to train Whisper. In general, we don't know what parts of the models are actually informing the observed correlation. The PPG model has millions of parameters to learn the mapping to PPG and then thousands of parameters to learn the classification from PPG feature vector to binary feature, and the Mel spectrogram classification also had thousands of parameters. The number of parameters might account for this disparity between accuracies and correlations, however, we don't have sufficient evidence to make that claim from the findings in this research.

The correlations provide interesting comparisons between methods and shine some light onto what these models are learning. One potential concern with the construction of our PPG feature vector is that it could learn more about duration (more frames in a summation is a larger number). If this were the case we would expect there to be a higher correlation between the PPGs and the speech rate. However, Figure 4.2 shows a low correlation between the two methods, thus, we suspect that the PPGs are using more phonetic information than length information in the modelling.

5.1 Limitations

This study has several important limitations that should be considered while discussing and interpreting the results.

Firstly, our data that was used to rank speakers, as well as train the classifiers was not ideal in a few ways. Our data from speechocean is comprised of short L2 utterances, this means that, in this context, accent can only be learned over a short context window. Short context windows make accent and speaker identification tasks

difficult (Chakroun et al., 2018). Accent may be better learned over a longer utterance where there is more linguistic and acoustic differences. For example, it may be difficult to discriminate an accent between a Spanish and Portuguese speaker if they read a short sentence in English, but a longer sentence would have more information about the speaker identity and accent. This research only used data from Mandarin speakers of English, which is an obvious limitation that it only understands the correlations and accuracies of this language pair. Another data limitation is that we didn't have large corpora of Mandarin L2 speech, so we had to train with small data sets.

Secondly, there were limitations on the models used in the training of classifiers. The models used in this study were relatively small, primarily due to the constrained size of the dataset. This size limitation likely prevented the models from capturing the full complexity of the tasks. We hypothesize that the use of larger models, necessarily coupled with more L2-accented speech data, would yield improved results. Additionally, a comprehensive hyper-parameter sweep was not conducted, leaving open the possibility that the models were not optimally tuned. The fact that the models, during training, did not converge to a loss of zero suggests that they may have been underfitted or not complex enough to fully capture the patterns in the data.

Thirdly, there were several limitations in the data collection and surveying processes. One significant challenge was getting survey participants to disentangle accent from fluency. This is a difficult listening task, even for expert listeners. The entanglement of accent and fluency may have introduced ambiguity in the data and affected participant's ability to accurately rank people on accentedness. We note that there was no training about the differences between accent and fluency, because we didn't want to interfere with people's initial impressions of accent strength. Furthermore, the study did not control for the diversity of accent backgrounds among participants in the preliminary round. This trade off exchanged potential variability in the results for more participants and results. We would have liked to control this variable so that all survey participants were of the exact same linguistic and accent background. Additionally, while the participants were all native English speakers, the study did not differentiate between monolingual and bilingual individuals. Because monolinguals and bilinguals perceive language differently (Astheimer et al., 2016), this could have impacted the results and subsequent analysis. Finally, during the final survey collection on Prolific, there was no mechanism in place to identify and exclude bad actors during data collection, potentially compromising the integrity of the dataset and introducing noise that could affect the study's findings.

Lastly, we would like to note that there is a large bias in research concerning the presence of English in accent studies. This is unfortunately reflective of data availability, however, we hope that this research can motivate the collection of diverse data as well as help generalise accent intensity learning to other language pairs, including non-English speakers.

5.2 Future Work

There are several avenues for future research that could build on the findings of this study. There is a lot of research left to do before speech synthesis models understand L2 speech and its variability.

One direction of research involves gaining a deeper understanding of why certain models outperform others in the task of accent strength identification and assessment. This research identifies that Whisper is better than PPGs and PPGs are better than Mel spectrogram classification, but understanding why is a significant and interesting question. What does Whisper learn that makes it better performing? There are too many confounding variables in this research to be able to answer this question.

Another direction for future work involves enhancing the modelling techniques used. As stated in the previous section, there was no hyper-parameter tuning on the models. Better modelling could also involve exploring more sophisticated approaches to modelling the influence of a speaker's first language (L1) on their second language (L2) speech. A promising approach worth investigating is the use of a “*Listen, Attend, Identify*” framework used for NLI, where a softmax classifier is trained to identify a speaker's L1 from a set of possible languages (Ubale et al., 2018). More powerful models could be utilised to understand the complex relationship between a speaker's L1 and their accented L2 speech.

Further research is also needed to examine whether the correlations observed in this study hold across different language pairs, beyond the Mandarin-English pairing in this study. Understanding the relationship between language pairs in speech in a broader linguistic and acoustic context would be crucial for developing more generalise multi-lingual accent strength models.

Finally, future work should also use speech synthesis models that use these findings about perception of accent strength to control accent strength in speech synthesis. These synthesised speech samples would be important to assess whether these variations are perceived as consistent and natural by listeners. This line of research could

lead to significant advancements in how we understand, measure, and control accent strength in a variety of linguistic and cultural contexts. All this future research would lead us to better accent training and representation of natural variability in speech, particularly L2 speech.

Chapter 6

Research Summary

Accented speech is ubiquitous in our lives and in our datasets. The degree of accent strength is a dimension of control that current works have not been able to manage successfully. There are no datasets, that we are aware of, that label L2 accent intensity either as discrete labels or continuous values. Accent strength is hard to control because it is difficult to measure. In an effort to measure accent strength, this research augments an entire corpus of Mandarin Chinese L2 speakers of English and ranks them in terms of their accent strength. This augmentation creates a new novel dataset called *ContinuousAccent*. This ranking and subsequent scoring can serve as a basis of understanding to help improve both future modelling and evaluation of accent intensity. This research discovers that Whisper language identification is the strongest correlation to human perception of accent strength in both continuous and discrete spaces compared to several other approaches taken. More work should be done to explicitly model accent strength, and expand to multiple language pairs. By advancing both the theoretical frameworks and practical tools for accent modeling, we seek to build inclusive and accurate speech technologies that better reflect the diversity of modern global speech expression.

Bibliography

- Acheme, D. E., & Cionea, I. A. (2022). “oh, i like your accent”: Perceptions and evaluations of standard and non-standard accented english speakers. *Communication Reports*, 35(2), 92–105.
- Aksënova, A., Chen, Z., Chiu, C.-C., van Esch, D., Golik, P., Han, W., King, L., Ramabhadran, B., Rosenberg, A., Schwartz, S., et al. (2022). Accented speech recognition: Benchmarking, pre-training, and diverse data. *arXiv preprint arXiv:2205.08014*.
- Aksënova, A., van Esch, D., Flynn, J., & Golik, P. (2021). How might we create better benchmarks for speech recognition? *Proceedings of the 1st workshop on benchmarking: Past, present and future*, 22–34.
- Astheimer, L. B., Berkes, M., & Bialystok, E. (2016). Differential allocation of attention during speech perception in monolingual and bilingual listeners. *Language, Cognition and Neuroscience*, 31(2), 196–205.
- Baas, M., van Niekerk, B., & Kamper, H. (2023). Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*.
- Bhela, B. (1999). Native language interference in learning a second language: Exploratory case studies of native language interference with target language usage.
- Biadsky, F., Weiss, R. J., Moreno, P. J., Kanevsky, D., & Jia, Y. (2019). Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *arXiv preprint arXiv:1904.04169*.
- Chakroun, R., Frikha, M., & Beltaifa zouari, L. (2018). New approach for short utterance speaker identification. *IET Signal Processing*, 12(7), 873–880.
- Chen, M., Chen, M., Liang, S., Ma, J., Chen, L., Wang, S., & Xiao, J. (2019). Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding. *Interspeech*, 2105–2109.
- Churchwell, C., Morrison, M., & Pardo, B. (2024). High-fidelity neural phonetic posteriorgrams. *arXiv preprint arXiv:2402.17735*.

- Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in psychology*, 3, 479.
- Cumbal, R., Moell, B., Lopes, J., & Engwall, O. (2024). You don't understand me!: Comparing asr results for l1 and l2 speakers of swedish. *arXiv preprint arXiv:2405.13379*.
- Cutler, C. (2014). Accentedness, "passing" and crossing. *Social dynamics in second language accent*, 145–167.
- Derakhshan, A., & Karimi, E. (2015). The interference of first language and second language acquisition. *Theory and Practice in language studies*, 5(10), 2112.
- Ding, S., Zhao, G., & Gutierrez-Osuna, R. (2022). Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech & Language*, 72, 101302.
- Dragojevic, M., Berglund, C., & Blauvelt, T. K. (2018). Figuring out who's who: The role of social categorization in the language attitudes process. *Journal of Language and Social Psychology*, 37(1), 28–50.
- Gass, S. M., Mackey, A., & Pica, T. (1998). The role of input and interaction in second language acquisition: Introduction to the special issue. *Modern Language Journal*, 299–307.
- Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the united states. *Journal of language and social psychology*, 29(2), 224–234.
- Godoy, E., Rosec, O., & Chonavel, T. (2009). Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling. *InterSpeech 09: 10th Annual Conference of the International Speech Communication Association*.
- Graham, C. (2021). L1 identification from l2 speech using neural spectrogram analysis. *Interspeech*, 2021, 3959–3963.
- Herbrich, R., Minka, T., & Graepel, T. (2007). Trueskill(tm): A bayesian skill rating system (Advances in Neural Information Processing Systems 20). *Advances in Neural Information Processing Systems 20*, 569–576. <https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system/>
- Humayun, M. A., Yassin, H., & Abas, P. E. (2022). Native language identification for indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions. *Speech Communication*, 139, 92–101.

- Jesney, K. (2004). The use of global foreign accent rating in studies of L2 acquisition. *Calgary, AB: University of Calgary Language Research Centre Reports*, 1–44.
- Kaners, S., Cucchiarini, C., & Strik, H. (2009). The goodness of pronunciation algorithm: A detailed performance study.
- Knight, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314), 436–439.
- Kominek, J., & Black, A. W. (2004). The cmu arctic speech databases. *Fifth ISCA workshop on speech synthesis*.
- Kukk, K., & Alumäe, T. (2022). Improving language identification of accented speech. *arXiv preprint arXiv:2203.16972*.
- Lecumberri, M. L. G., Barra-Chicote, R., Perez, R. R., Yamagishi, J., & Cooke, M. (2014). Generating segmental foreign accent. *INTERSPEECH 2014 15th Annual Conference of the International Speech Communication Association*, 1302–1306.
- Lesnichaia, M., Mikhailava, V., Bogach, N., Lezhenin, Y., Blake, J., & Pyshkin, E. (2022). Classification of accented English using CNN model trained on amplitude mel-spectrograms. *INTERSPEECH*, 3669–3673.
- Liu, D., Xu, J., Zhang, P., & Yan, Y. (2021). A unified system for multilingual speech recognition and language identification. *Speech Communication*, 127, 17–28.
- Liu, R., Sisman, B., Gao, G., & Li, H. (2024). Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2188–2201. <https://doi.org/10.1109/TASLP.2024.3378110>
- Liu, S., Wang, D., Cao, Y., Sun, L., Wu, X., Kang, S., Wu, Z., Liu, X., Su, D., Yu, D., et al. (2020). End-to-end accent conversion without using native utterances. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6289–6293.
- Lyth, D., & King, S. (2024). Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.
- Ma, L., Zhang, Y., Zhu, X., Lei, Y., Ning, Z., Zhu, P., & Xie, L. (2023). Accent-vits: Accent transfer for end-to-end tts. *National Conference on Man-Machine Speech Communication*, 203–214.
- Markl, N., & Lai, C. (2023). Everyone has an accent. *Interspeech 2023*, 4424–4427.
- Marx, N. (2002). Never quite a 'native speaker': Accent and identity in the L2-and the L1. *Canadian Modern Language Review*, 59(2), 264–281.

- Melechovsky, J., Mehrish, A., Sisman, B., & Herremans, D. (2022). Accented text-to-speech synthesis with a conditional variational autoencoder. *arXiv preprint arXiv:2211.03316*.
- Mohammadi, S. H. (2015). Reducing one-to-many problem in voice conversion by equalizing the formant locations using dynamic frequency warping. *arXiv preprint arXiv:1510.04205*.
- Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48(2), 159–182.
- Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of l2 speech. *Studies in second language acquisition*, 28(1), 111–131.
- Ordin, M., & Polyanskaya, L. (2015). Perception of speech rhythm in second language: The case of rhythmically similar l1 and l2. *Frontiers in psychology*, 6, 316.
- Peri, R., Li, H., Somandepalli, K., Jati, A., & Narayanan, S. (2020). An empirical analysis of information encoded in disentangled neural speaker representations. *arXiv preprint arXiv:2002.03520*.
- Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an l2: A review. *Journal of phonetics*, 29(2), 191–215.
- Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, 58, 1–21.
- Qian, Y., Evanini, K., Wang, X., Suendermann-Oeft, D., Pugh, R. A., Lange, P. L., Molloy, H. R., & Soong, F. K. (2017). Improving sub-phone modeling for better native language identification with non-native english speech. *Interspeech*, 2586–2590.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning*, 28492–28518.
- Saito, K., & Hanzawa, K. (2018). The role of input in second language oral ability development in foreign language classrooms: A longitudinal study. *Language Teaching Research*, 22(4), 398–417.
- Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5), 336–347.

- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591–611.
- Shimizu, R., Yamamoto, R., Kawamura, M., Shirahata, Y., Doi, H., Komatsu, T., & Tachibana, K. (2024). Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12672–12676.
- Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 132–157.
- Tännander, C., O'Regan, J., House, D., Edlund, J., & Beskow, J. (2024). Prosodic characteristics of english-accented swedish neural tts. *Speech Prosody 2024*, 1035–1039.
- Ubale, R., Qian, Y., & Evanini, K. (2018). Exploring end-to-end attention-based neural networks for native language identification. *2018 IEEE spoken language technology workshop (SLT)*, 84–91.
- Weigel, A. P., Liniger, M., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(630), 241–260.
- Wells, J. C. (1982). *Accents of english: Volume 1* (Vol. 1). Cambridge University Press.
- Wieling, M., Veenstra, J., Adank, P., & Tiede, M. (2017). Articulatory differences between l1 and l2 speakers of english. *The 11th International Seminar on Speech Production*.
- Williams, S., Foulkes, P., & Hughes, V. (2024). Analysis of forced aligner performance on l2 english speech. *Speech Communication*, 158, 103042.
- Ye, H., & Young, S. (2006). Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), 1301–1312.
- Youngsun, C., & Hosung, N. (2021). A comparison of l1 and l2 speech phonetic posteriorgrams for applications in pronunciation training. *Academic Journal of Foreign Language Education Research*, 35(1), 293–304.
- Yu Chen, J. H., & Zhang, X. (2019). Sell-corpus: An open source multiple accented chinese-english speech corpus for l2 english learning assessment. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

- Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., Li, K., Povey, D., & Wang, Y. (2021). Speechocean762: An open-source non-native english speech corpus for pronunciation assessment. *arXiv preprint arXiv:2104.01378*.
- Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A., & Ramabhadran, B. (2019). Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.
- Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. (2023). Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.
- Zhao, G., Ding, S., & Gutierrez-Osuna, R. (2019). Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. *Interspeech*, 2843–2847.
- Zuluaga-Gomez, J., Ahmed, S., Visockas, D., & Subakan, C. (2023). Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *arXiv preprint arXiv:2305.18283*.

Appendix A

Additional Materials

A.1 Qualtrics




You will be asked questions where you are to rank 3 sample of speech according to perceived accent strength.

Every sample comes from a person who speaks English as a second language. Mandarin Chinese is their first language.

Accent strength is not very well defined, but do your best and follow your gut. Thank you for your help.

Next page >

Figure A.1: These were the instructions given to all participants of both the preliminary and final survey. All participants were native English speakers.



Order these utterances in terms of accentedness. 1 (best/most native) 3 (worst/least native).

1: ▶ 0:00 / 0:04 ⏮ ⏭ ⋮

2: ▶ 0:00 / 0:04 ⏮ ⏭ ⋮

3: ▶ 0:00 / 0:02 ⏮ ⏭ ⋮

[Next page >](#)

Figure A.2: This is an example ranking question. The participant listened to each audio and dragged and dropped the audio into the 1st, 2nd, or 3rd place depending on how they would rank the three speakers in terms of accent strength. There were 25 questions in the preliminary survey and 50 questions in the final Prolific survey.

Welcome to the research study!

We are interested in understanding Accentedness in Speech. You will be presented with speech utterances relevant to accent and asked to answer some questions about it. Please be assured that your responses will be kept completely confidential.

The study should take you around 30 minutes to complete, and you will receive £5 for your participation. Your participation in this research is voluntary. You have the right to withdraw at any point during the study, for any reason, and without any prejudice. If you would like to contact the Principal Investigator in the study to discuss this research, please e-mail s2598458@ed.ac.uk.

By clicking the button below, you acknowledge that your participation in the study is voluntary, you are 18 years of age, and that you are aware that you may choose to terminate your participation in the study at any time and for any reason.

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

☐ I consent, begin the study

☐ I do not consent, I do not wish to participate

[Next page >](#)

Figure A.3: This is the consent form that was presented to the participants of the final Prolific survey. This information was all passed through an ethics committee who approved the payment and conditions of the experiment.

A.2 Qualtrics Results

Summary of QID16: Order these utterances in terms of accentedness. 1 (best/most native) 3 (worst/least native).

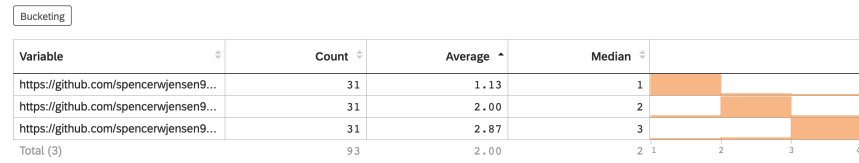


Figure A.4: This is an example of the results that were returned by Qualtrics. Question 16 had 31 responses and the distribution of those 31 responses is seen to the right. Question 16 was the control question with the highest inter-rater agreement.