Evaluating Automatic Spoken Fluency Assessment Methods in Spoken English

Sachien Fernando

Advisor: Robert Frank

Submitted to the faculty of the Department of Computing and Language in partial fulfillment of the requirements for the degree of Bachelor of Arts

Department of Computational Linguistics
Yale University

DATE SUBMITTED

[Revised May 6, 2024]

Abstract

In recent years, there has been an increase in the use of automated methods to evaluate the spoken proficiency of non-native English speakers. The relationship between perceived fluency and a speaker's perceived accentedness is well-researched for human judges, and English is strongly affected due to the high inter-regional variation between speakers. However, no previous studies have explored the impact of such variations in the native language on automated English fluency ratings. This paper explores the effects of the similarity of a speaker's native language to English and its representation within the training data on the fairness of their automated fluency prediction. The paper uses the Rated L2 Speech Corpus and the Speech Accent Archive to train an automated fluency scoring model. We conducted several experiments to judge the best features and models for these purposes and to ensure that the two datasets could mix within the model. Additionally, the speaker's age of language acquisition was used as a proxy for fluency, and the correlation between the age of onset and the score for a language was viewed a as a measure of the fairness of judgement. Second, the trained model was tested on subsets of L2 English speech from the Speech Accent Archive based on the native language. Results suggest that the similarity of the native language to English and the presence of native language samples in the training data both impact automated fluency judgments, with the latter having a stronger effect on the fairness of fluency judgement.

Table of contents

1. Introduction

As of 2023, English is the world's most widely spoken language, with over 1.5 Billion speakers in 58 countries and 28 non-official entities. (Eberhard et al. 2023 : 1). A combination of regional influences and naturally occurring linguistic variation cause different speakers of English to acquire different target dialects. Due to this, the native English spoken inter-regionally varies especially greatly in both accentedness and morphosyntactic properties. These issues challenge the reliability of spoken English proficiency judgment, exacerbating concerns surrounding the increasing adoption of automated methods of scoring spoken fluency. Partially or entirely automated assessments like the TOEFL iBT examination and the Duolingo English Test (DET) are often accepted or even required for non-native English speakers to prove their spoken proficiency. However, machine-learning models tend to internalize systemic inequities within their training data, and models used for Automatic Speech Recognition have also shown higher Word Error Rates for racial and gender minorities and speakers of nonstandard varieties of English (Langenkamp et al. 2020; O'Donnell 2019; Ngueajio and Washington 2022; Chan et al. 2022). These concerns raise the question — how might similar biases affect automated spoken fluency prediction?

This paper investigates the effect of a non-native (L2) English speaker's native language (L1) on an automated assessment of their fluency. Specifically, it explores how a speaker's native language affects the assessment based on its similarity to English and its presence in training data. I created an automated fluency assessment model based on techniques reported in the TOEFL SpeechRater and related research. The training datasets used are the Rated L2 Speech Corpus (RLSC) and sections of the Speech Accent Archive (SAA). While rated, the L2 Rated Speech Corpus was too small and did not contain native speakers. Conversely, the Speech Accent Archive did not contain any fluency ratings. They also had

different modalities, consisting of spontaneous and read speech respectively. I had to conduct several checks on the data to answer the question reliably without any unwanted confounds interfering. Experiment 1 ensures that a narrow fluency description is sufficient for the Speech Accent Archive. Experiment 2 shows that the neural network model can generalize from L2 spontaneous speech to native Read speech. Experiment 3 shows that a speaker's age of English onset can serve as a proxy for their fluency. Finally, Experiment 4 uses the model trained on native SAA and RLSC data to test the fairness of fluency assessment across native Spanish, German, Korean, and Farsi speakers.

2. Background

2.1 Defining Fluency

Although the various aspects of spoken fluency are well-researched, they have all assumed different definitions and scopes depending on appropriate contextual factors. Additionally, the term has been used interchangeably with 'proficiency.' Therefore, to evaluate language fluency, one must choose a scope and definition that fits their particular case and disambiguate this from the others.

Most contemporary characterizations of proficiency base themselves on the 'Complexity, Accuracy, and Fluency' (CAF) framework. The elements can be broken down into subcategories, but they are broadly defined as follows: Complexity refers to "the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2", accuracy is "the ability to produce target-like and error-free language", while fluency is "the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation" (Housen et al. 2012: 2). However, this definition of fluency is relatively recent and has a narrower scope than other definitions. In less recent scholarship, there have been two general senses of fluency. The broad definition serves as a "cover term for oral proficiency." The narrower scope views fluency, like within the CAF framework, as motivated by the idea that a learner may be "fluent but grammatically inaccurate," "fluent but lack(s) a wide and varied vocabulary", or "speak correctly but not very fluently"(Lennon 1990: 390). In this way, the narrow scope cleaves general oral proficiency into three independent parts, of which fluency is one.

The metrics used to judge a speaker's fluency are highly influenced by the modality and context of testing. During a reading test, the complexity of the elicitation text cannot influence fluency judgments. Meanwhile, during a spontaneous or conversation-based test, a

speaker might be expected to use vocabulary appropriate to their context and environment. Additionally, while they can be judged on accuracy and fluency, studies have shown that spontaneous and read speech "are significantly different acoustically as well as linguistically." (Nakamura et al. 2008: 171). However, the literature maintains that there is "no clear dividing line" even though "intonation, phoneme duration, and spectral features all contain cues" to distinguish between the two (Laan 1992: 64).

The remainder of this paper and the following experiments will use the narrow scope of *fluency*. This choice accounts for spontaneous and read speech in the training and testing data since the narrow scope also measures certain aspects of fluency in non-spontaneous speech. Experiment 1 will also include "features based on properties of the speech file's transcription in some instances, which involves assessing a broader version of fluency more suited to exclusively spontaneous speech. However, this is merely to be used as a baseline to check whether purely speech-based features are sufficiently predictive. [1]

2.2 Fluency Judgements and Algorithmic Bias

Humans are far from objective when being judges of fluency. Research has shown that factors such as familiarity with the relevant accent influence a speaker's "perceived intelligibility while actual comprehensibility seems to be less affected by this aspect" (Beinhoff 2014: 58). Therefore, individual human fluency judgments appear to be susceptible to the judge's familiarity with a speaker's accent. "Oral fluency and foreign accent" are also considered two important factors in the task of "distinguish[ing] L2 from L1 speech production" (Pinget et al. 2014: 349). In fact, at higher speaking rates, the heavy presence of an unfamiliar accent causes a "decrease in comprehension" with human judges (Anderson-Hsieh and Koehler 1988: 562). These findings are counterintuitive to the well-established idea that high speaking rates should be an important indicator of spoken fluency. With humans being biased with fluency predictions, as we tend to be in other aspects of our lives, one might hope that cold, unfeeling

[1] See Appendix A for more information about feature types and their formation.

computer programs can provide a more objective metric of spoken fluency.

Until recently, expert human scorers have been utilized as the primary evaluation source for spoken English fluency. However, they often prove to be "costly, time consuming, and subject to … rater fatigue, rater bias etc." Recent advances, including the "use of Deep Neural Network algorithms for training automatic speech recognition models" mean that machine learning models "can be used productively in a wide range of services" (Zechner and Evanini 2019: 3). To that effect, many institutions and companies have developed proprietary models to score non-native English speaker fluency. These tests are often taken by English learners to validate and attest their proficiency in both professional and academic context. The vast majority of these test takers live in countries where English is not spoken natively, and most test takers are non-native speakers themselves (Cardwell et al. 2022: 24).

Industry models such as the Duolingo English Test (DET) and the ETS SpeechRater claim to take measures towards achieving equity in their tests. For instance, DET items undergo human Fairness and Bias reviews to identify sources of "construct- irrelevant variance" and "potentially problematic items" to "ensure items are fair towards test takers of diverse identities and backgrounds" (Cardwell et al. 2022: 16). Similarly, the SpeechRater Acoustic Model aims to tackle this issue in the training process, through data "collected from non-native English test takers with diverse L1s" (Loukina et al. 2017: 4). In either case; no mention is made of explicit efforts to account for the potential variance in their test takers. These efforts toward equity might make progress towards ensuring the test items are more fair. However, these measures would not affect any inherent inequality in the mechanism of testing itself. Consequently, it remains unclear whether the model would be able to accurately and fairly evaluate spoken English from across the world according to their respective fluency levels. Due to their proprietary nature, most specific architectural details of the models are not publicly available. However, it is known that models such as ETS SpeechRater and the Duolingo English Test use a combination of acoustic features and features derived from Automatic Speech Recognition (ASR) as values into the neural network model. For the

SpeechRater, features are divided into categories and subcategories. The main categories are Delivery, Language Use, and Topical Development. These categories are intended to map to the common linguistic assessment features of Fluency and Pronunciation, Vocabulary and Grammar, and Content and Discourse Coherence (Zechner and Evanini 2019: 99). Duolingo appears to be more opaque with their scoring methods. What is known is that OpenAI's Whisper model is utilized for transcription, and that "transcription-based and acoustic features extracted from oral responses using computational methods" are given weighted scores based on their assumed "importance in the construct of speaking" (Park et al. 2023: 11).

Across all categories, the ETS SpeechRater collects 110 features as input for their automated scoring system. However, in the official results, the SpeechRater only presents 7 main categories. These include Speaking Rate, Sustained Speech, Pause Frequency, Repetitions, Rhythm, Vowels, and Vocabulary Depth. [2] The features were selected through a combination of "initial test[ing]", "psychometric evaluation", "ability to cover a range of language phenomena", "amenab[ility] to measurement", and "correlations with human score" (Zechner and Evanini 2019: 162). These individual features are represented on a percentile Scale. For the actual scoring, the model uses a simple Multiple Linear Regression model. On testing, the Linear regression model had the lowest r and RMSE values, when compared with other techniques like Random Forest, Support Vector Machines, and Elastic Net machine learning techniques. [3] It is important to remember that most of these features are subject to large inter-regional variation. For example, British and American English - arguably the two dominant varieties of English - exhibit significantly divergent stress patterns (Berg 1999). Additionally, research also suggests differences in vowel duration within American, British, and New Zealand English (Cho 2016). Therefore, even native English speakers with different stress patterns could be penalized through the inclusion of these features.

[2]Details about the quantitative nature of these features are explained in the Feature Extraction section of Methods.

[3]RMSE stands for Root Mean Squared Error, while r values refer to the Pearson Correlation Coefficient.

3. Experiment 1: Capturing Fluency

Read speech lends itself to a narrow definition of oral proficiency. Meanwhile, one can evaluate spontaneous speech for its Complexity, Accuracy, and Fluency. The first step one must take before combining read and spontaneous speech datasets is that they should be able to be evaluated along the same conception of fluency. One way to assess both read and spontaneous speech along the same fluency scale is to utilize the subset of the scale common to both members. However, one must first ensure that the narrower subset of features is sufficiently predictive of the overall fluency score assigned to the spontaneous speech. In practice, this involves ensuring that a model trained on input features evaluating narrow fluency performs comparably to a model trained on all available features.

The following experiment is an in-domain test on the Rated L2 Speech Speech Corpus. The experiment will use cross-validation to divide the test set into five parts. Therefore, 80 percent of the L2 speech data will train the model for each of the five iterations, while the remaining 20 percent will be used for testing.

3.1 Dataset

The dataset used in Experiment 1 is the Rated L2 Speech Corpus. The Rated L2 Speech Corpus is a "rated database of spontaneous speech produced by [28] second language (L2) learners of English", developed at the University of Illinois (Yoon et al. 2009: 662). The speakers were all adults, with an average age of 27.7 and a native language breakdown described in Table 3.1. The data consisted of audio recordings of spontaneous responses to questions based on "formats similar to that of the TOEFL iBT." Each recording also had a corresponding fluency rating, which was "assessed by two phonetically trained, experienced ESL instructors" (Yoon et al. 2009: 662). The ratings ranged over a scale of 0-4, where

Native Language	Number
Korean	14
Chinese	8
Spanish	3
Arabic	1
French	1
Thai	1

Table 3.1: Native Language Breakdown of L2 Rated Speech Corpus

0 represented "no attempt to respond or … unrelated responses", while four represented "native-like fluency." The rubric adopted for assessment was that of the iBT TOEFL examination (See Appendix B). Figure 3.1 describes the probability distribution of the scores, along with a Normal Distribution curve approximated over it using the distribution's mean and standard deviation.

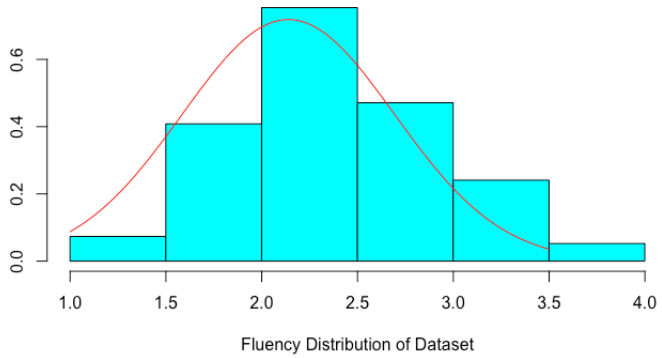

Figure 3.1: Probability Distribution of Fluency Ratings in Rated L2 Speech Corpus

The grades were relatively normally distributed and had a mean of 2.67 and a Standard Deviation of 0.55. Therefore, the data provides a reasonably diverse sample of non-native English speakers at different stages of assessed fluency.

3.2 Methods

3.2.1 Features

Experiment 1 utilizes three feature sets: acoustic, Phonetic, and Transcription-Based. They are described below. Appendix A contains the complete list of features used.

- **Acoustic Features:** Acoustic features capture the audio file's voice quality properties. They refer to the "acoustic characteristics of voice signals." These features include Jitter and Shimmer, which are "quantified as the cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively, "spectral flux, which measures the spectrum's variability over time, and the cepstral coefficients. The features include the means and standard deviations over the relevant portions of the audio file. Overall, there are 14 acoustic features. While these features do not directly correspond to any conceptions of fluency, they have been used previously as features in various speech models for the oral reading proficiency of L2 English children (Kim et al. 2021). All of the acoustic features were extracted using the Python package *opensmile*. Initial tests were conducted with a larger set of features, and the features with the highest coefficients in the prediction were used in the final model.
- **Phonetic Features:** Phonetic features represent higher-level speech features and are more linguistically motivated. These include the Speaking rates measured by words and syllables, number of pauses per second, average speech chunk length, and average voiced and unvoiced syllable length. Overall, there are nine features in this set. The speech rate in words were taken from the transcription, normalized by time. The speech rates in syllables were calculated using python package *syllapy*. Voiced and unvoiced syllables were calculated using *opensmile*. A chunk of speech is measured as the length of time between pauses. Therefore, the average chunk of speech features is the average length of a chunk, normalized by the number of words.

- **Transcription-Based Features:** These features are associated with the English transcription of the speech file. They include measures of readability, complexity, and average sentence structure depth. Transcription-based features measure the Complexity and Accuracy dimensions of general oral proficiency. Therefore, they can measure spontaneous fluency, but do not apply to read speech fluency.

Official tests like the SpeechRater and the DET often use Phonetic and transcription-based features since they usually measure spontaneous fluency. Meanwhile, acoustic features have historically been used in research for reading proficiency judgment. However, nothing prevents them from being used to augment spontaneous fluency prediction. Therefore, this experiment will combine acoustic features with the other features used for spontaneous speech to help the model when being ablated.

3.2.2 Models

When creating a fluency prediction model, we must choose the model to use. The SpeechRater model uses a Linear regression model but tests a variety of others, including the Random Forest and Support Vector Machine Regression Techniques (Zechner and Evanini 2019). Linear Regression, Random Forest, and Multilayer Perceptron models are all tested in this experiment. The models are detailed as follows:

- **Linear Regression:** This implementation uses 'Least Squares' regression to find the optimal coefficients to assign to each input parameter to minimize the Mean Squared Error between predicted and target values. No parameters were needed to be modified.
- **Random Forest:** A random forest returns a "collection of decision trees that have been trained on randomly selected subsets of the training instances and explanatory variables." The model returns the mean value predicted by their constituent trees. No parameters were needed to be modified.
- **Multilayer Perceptron:** The MLP consists of two hidden layers with 10 and 8 neurons respectively with 'relu' activation. Between each layer, dropout is conducted

to avoid overfitting at a rate of 0.2. Finally, the output layer consists of a single linear activation function. This model was trained for 100 epochs.

3.2.3 Experimental Setup

Each of the models listed above will run two separate tests. First, all possible features are used to train the model. During the second run, the transcription features will be ablated. Since the test is an in-domain on the Rated L2 Speech Corpus, we will use 5-fold cross-validation, which means the dataset is divided into five sections of roughly equal size. The model will be trained on 80 percent of the data for each run, while 20 percent is held out for testing. This way, every data point is tested once and used as training data 4 times. Table 3.1 shows the number of input features with and without ablation.

Feature Label	Input Dimension Size
All Features	26
Ablated Features	23

Table 3.2: Input Dimension Size by Feature Label

The experiment's ultimate goal is to measure the model's predictive power within the dataset - with and without the transcription features. The Root Mean Squared Error (RMSE) and the r^2 values usually measure the model's fit. The former measures the difference between the predicted and target values, while the latter measures the variability in the original data that the model predicts. Therefore, an ideal model would have an RMSE of 0 and a r^2 value of 1. However, we are less interested in the actual RMSE and r^2 values for this experiment. Instead, we want to compare the differences in scores of the ablated and non-ablated models.

3.3 Results

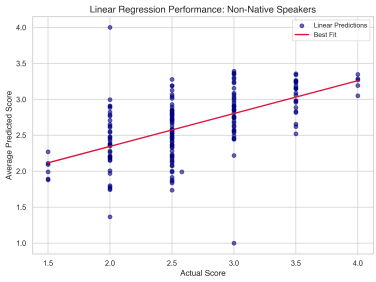

Figure 3.2: Linear with All Features

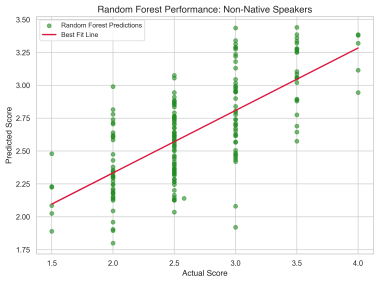

Figure 3.3: Random Forest with All Features

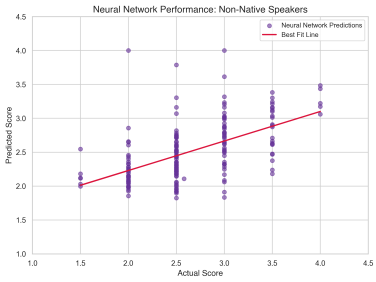

Figure 3.4: Neural with All Features

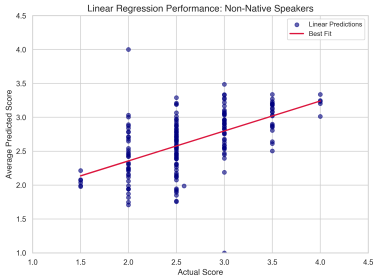

Figure 3.5: Linear with Ablated Features

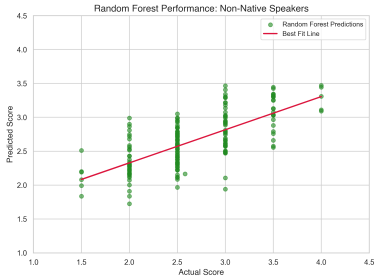

Figure 3.6: Random with Ablated Features

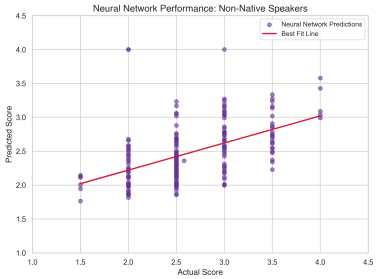

Figure 3.7: Neural with Ablated Features

Model	**RMSE**	R^2	**L2 Mean**	**L2 StdDev**
Linear Regression	0.4717	0.2616	2.6399	0.4435
Random Forest	0.3968	0.4774	2.6399	0.3771
Neural Network	0.5123	0.1291	2.5357	0.4344

Table 3.3: Values for All Features

Model	**RMSE**	R^2	**L2 Mean**	**L2 StdDev**
Linear Regression	0.4751	0.2509	2.6416	0.4368
Random Forest	0.3957	0.4804	2.6428	0.3871
Neural Network	0.5395	0.0980	2.4494	0.4537

Table 3.4: Values for Ablated Features

In the following graphs, the assigned scores from the Rated L2 Speech Corpus are on the x-axis, while the predicted scores are on the y-axis. Figures 3.2 - 3.5 represent the respective model with all features, with its values in Table 3.3. Figures 3.5 - 3.7 represent the ablated model, with the values found in Table 3.4.

Within the models containing all features, the Random Forest appears to have the best performance, having the lowest RMSE and highest r^2 values. Still, the models are comparable in performance, with the Neural Network performing worse than the regression models. The ablated models show a similar pattern, with all models having similar values to the original dataset's mean of 2.67. Significantly, each model can predict higher scores for more fluent speakers and lower scores for less fluent ones. The final crucial observation is that we find highly similar results across the ablated and non-ablated models when comparing the models. The p-value of the Linear Regression model was 0.806, the Random Forest was 0.499, and the Neural Network was 0.012. Therefore, the only model in which there was no statistical significance in the difference between the full feature set and the ablated feature set was in the Neural Network. [1]

3.4 Discussion

Our primary aim in this experiment was to compare the performance of different models in predicting fluency in L2 speech. While this was not the sole focus of our study, the data suggests that the Random Forest model performed the best, with the Linear Regression model also showing similar results. Interestingly, the regression models outperformed the Neural Network, indicating their potential usefulness in data-scarce scenarios. This corroborates the idea that regression models might be desirable in a highly data-scarce scenario.

Overall, the ablated models were nearly indistinguishable from the complete models, indicating that the audio data and their corresponding fluency metrics can be captured along a narrower notion of fluency compatible with read speech. As a result, the ablated

[1]The difference was considered to be statisically significant if the p-value was greater than 0.05

feature set will be used for all further experiments. Additionally, since the regression models performed similarly, the neural network could be underperforming due to insufficient data. However, the Neural Network model was the one in which the ablation caused the least effect. Since we could not ascertain for sure which model was the best, all the models, including the neural network, will continue to be used for Experiment 2.

4. Experiment 2: Can It Generalize?

Experiment 1 demonstrated how speech-based features are sufficiently predictive in a model tested on L2 Spontaneous Speech. However, the Rated L2 Speech Corpus has two problems associated with it. Firstly, it consists only of L2 Speakers. A typical fluency prediction test would have a large portion of their training data comprising speech from native speakers of English (Zechner and Evanini (2019) 2019). Secondly, the corpus is relatively small. It consists of data from only 28 speakers and 228 speech recordings. That much data, on its own, would neither represent practical models' data distribution nor sufficient data to train the models correctly. Therefore, we utilize a new dataset, the Speech Accent Archive, consisting of read speech (Weinberger and Kunath 2015). Therefore, before mixing the two within the same training data, we must ask whether a model trained on spontaneous speech can accurately predict the fluency of read speech.

4.1 Dataset

The new data introduced in this experiment comes from the native speaker subset of the Speech Accent Archive. George Mason University maintains the Speech Accent Archive, which consists of "native and non-native speakers of English all read[ing] the same English paragraph" (Weinberger and Kunath 2015). The primary purpose of this archive is to document various English accents spoken worldwide. The archive consists of 3031 recordings , each from a different speaker, with a total of over 390 native languages. The paragraph elicited from all the speakers is as follows:

> "Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

The above passage was chosen for elicitation as it "contains most of the consonants, vowels, and clusters of standard American English" (Weinberger and Kunath 2015).The Archive also collects various pieces of information from the speaker such as:

- Birthplace
- Other Languages Spoken
- Age of English Onset
- English Learning Method (Naturalistic/Academic etc.)
- Length of Residence in English Speaking Country

These elicitations are not rated for fluency and are not spontaneous. Therefore, if testing for fluency, one could not neatly fit predictions onto the TOEFL iBT rubric. Nevertheless, since native English speakers are fluent by definition, one could say that they automatically can be given a score of 4. There are 658 native speaker recordings in the SAA, with most speakers originally from regions like Australia, the United Kingdom, and the United States. However, there are also native speakers from other world regions, such as Trinidad, South Africa, India, and the United Arab Emirates. This regional diversity is helpful, as it likely mirrors the distribution of speakers in the training data of official fluency assessment models.

4.2 Methods

Experiment 1 showed that non-transcription features can reasonably predict the fluency of our corpus containing L2 Spontaneous Speech. Therefore, the experiment aims to test whether a model trained exclusively on non-transcription features of the Rated L2 Speech Corpus can generalize to predict the "native" status of the Native speakers in the SAA. Since

all models performed comparably, Experiment 2 reuses all the models from Experiment 1.

Experiment 2 is an out-of-domain test in two ways. The training data is the complete Rated L2 Speech corpus, which is spontaneous and consists entirely of L2 speakers. Meanwhile, the model tests the native speakers in SAA, which consists of read native speech. Therefore, it is out of domain by differing both in the read-spontaneous sense and the L2-native speaker sense. The models' capacities to generalize will be measured in a nuanced sense, not just by a single metric. One could expect that a model insensitive to the difference between read and spontaneous speech would judge any native speaker as more fluent than any non-native speaker. Ideally, all native speakers will be considered fluent and assigned a score of 4. A more realistic method of evaluating the score is to check whether the native speakers are, on average, assigned a higher score than the non-native speakers. Additionally, since transcription-based features are not in the model, the fluency prediction will not be affected by the actual content of the text elicitation.

4.3 Results

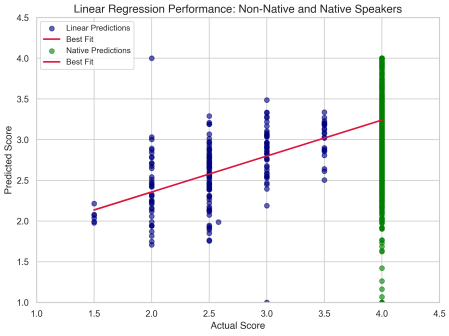

Figure 4.1: Linear Regression Prediction for Native Speakers

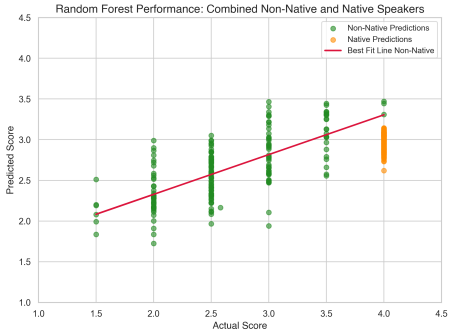

Figure 4.2: Random Forest Prediction for Native Speakers

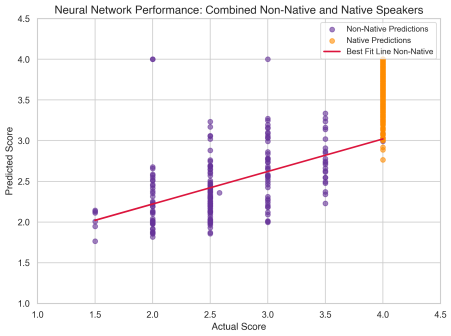

Figure 4.3: Neural Network Prediction for Native Speakers

Model	**3.5 Mean**	**3.5 Variance**	**Native Mean**	**Native Variance**
Linear Regression	3.048	0.054	2.955	0.393
Random Forest	3.128	0.063	3.001	0.008
Neural Network	2.803	101	3.877	0.022

Table 4.1: Mean and Variance of Model Values at 3.5 and Native

The data in the graphs are as follows. The x-axis consists of the actual or expected scores of the data points, while the y-axis shows the predicted score at each point. The data points on the graph consist of both the predicted scores of the training data - the same points from Experiment 1 — along with the native Read predictions. The training data points are distributed across the x-axis, while all the test data of this experiment are at the value of 4. The data being evaluated in this experiment is the prediction of native read fluency in relation to the L2 data from Experiment 1.

The models showed varying results in this experiment. Table 4.1 displays between the mean values and variances of both the predicted values in the Rated L2 Speech Corpus that have an actual rating of 3.5 and all the native scores. The Linear model displayed a variance in native values and a predicted mean value similar to that of the more fluent L2 speakers. Although the Random Forest had a minor variance, the predicted values were lower than those of the fluent L2 speakers. The Neural Network model was the only model with a reasonably low variance and a mean native speaker value higher than even the fluent L2 speaker predictions.

4.4 Discussion

Upon initial examination, Experiments 1 and 2 seem to present conflicting results. The first experiment indicated that the Random Forest's predicted values had the highest fit within the L2 spontaneous data, suggesting it could best generalize to predict native speech. However, its predictions for native speakers were lower than that of the non-native speakers with an actual score of 3.5. While the linear regressor managed to predict higher scores for native speakers, it was only the Neural network that managed to truly generalize from L2 Spontaneous speech to Native Read speech. It's worth noting that the similarity between the predicted scores and the actual scores is less important than the consistency of scores within the model. In other words, it's more beneficial for the prediction if the model clearly distinguishes between the predictions it gives speakers of different fluency levels than if the model's lowest and highest scores are closer to 1 and 4 respectively. A badly calibrated but internally consistent model is preferable to a well-calibrated but internally consistent one. Therefore, the Neural Network performs best when considering the generalization from L2 Spontaneous to Native Read. Since the other models show inconclusive results over the generalization, they will not be used in further experiments.

This experiment does not mix the Speech Accent Archive and the Rated L2 Speech Corpus. In Experiment 2, the training and test data differed along two dimensions: Read/Spontaneous and L2/Native. One could appeal to intuition: it makes sense that native speakers will score higher than L2 speakers, and the model picks up on this distinction. However, with the current information, one still needs to definitively rule out the possibility that the model is rating read speech as intrinsically higher and more fluent than spontaneous speech. Even though transcription-based features have been removed, there are documented distinctions between the acoustic properties of Read and Spontaneous speech, which could be causing this result. To distinguish between these potential causes, we would need to conduct more tests on the SAA.

5. Experiment 3: Finding a Proxy

The main drawback of the Speech Accent Archive is its absence of human-judged fluency ratings. Experiment 2 used the subset of the data that one could reasonably assume a score for - native speakers. However, we cannot assume the score of non-native English speakers without any examination. We should consider finding a heuristic for fluency to test the remainder of the SAA. Additionally, Experiment 2 did not clarify whether the model distinguished between read and spontaneous speech or L2 and native speech. We must answer these two questions before combining these datasets in our model. Although it does not contain fluency ratings, the SAA collects metadata from the recorders, which is explained fully in Section 4.1. Two valid data points are the "Age of English Onset" and the "Age" metrics. From there, we can find two hypothetical "proxies" for fluency - "Age of Onset" and "Years Learning English"[1]. They coincide with two intuitive ideas associated with fluency. On one hand, one might expect that a speaker's fluency has a positive correlation with the length of time they have spoken the language. One can conversely imagine that the earlier a speaker starts speaking English, the more fluent they are. Significantly, these two theories predict different correlations with fluency within them. A model accurately predicting fluency independently of the read-spontaneous distinction might follow one of these two patterns. There is another, more sobering, theory. A model trained exclusively on spontaneous speech may consistently rate all read speech as having high values. Experiment 3 will help us test our hypotheses.

[1]We can easily extract the "Years Learning English" metric by subtracting the "Age of English Onset" from the speaker's "Age"

5.1 Methods

Experiment 3 has a setup that is highly similar to that of the previous experiment. Since the Linear Regression and Random Forest models did not clearly mark native read speech as more fluent than L2 Spontaneous speech, we can assume that they could not generalize fluency predictions from spontaneous to read speech. Therefore, this experiment will only use the Neural Network model and the non-transcription feature bundle from Experiment 2.

These models, once trained, will be tested on native Spanish speakers from the SAA. The choice of native language for this experiment is arbitrary, since we would have expected some sort of distribution of scores for any native language. Since these speakers are L2 Speakers of English, we can expect them to have some distribution of actual fluency predictions. Therefore, if the models predict according to the fluency lines, they will also present their data along some distribution. If they score all read speech highly instead, the scores will be concentrated solely above three and be similar to the Native Read data from Experiment 2. These predictions will also be plotted against each individual's "Age of Onset" and their "Years Learning English," respectively, to see whether either metric correlates with the fluency predictions.

5.2 Results

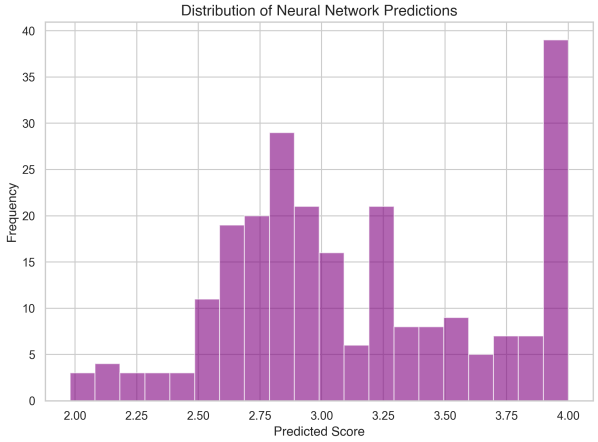

Figure 5.1: Histogram of Fluency Predictions for Spanish L2 Speakers

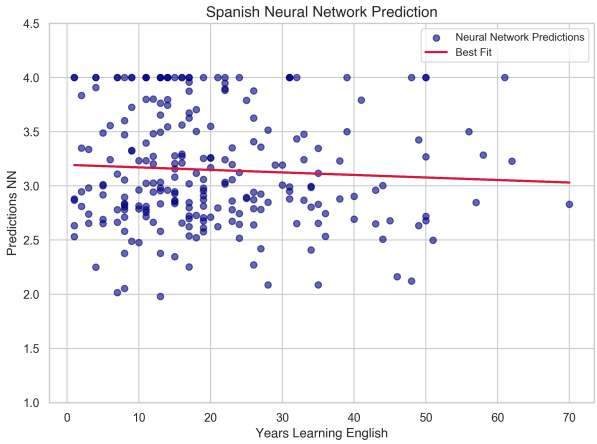

Figure 5.2: Spanish L2 Fluency Predictions against Years Spoken English

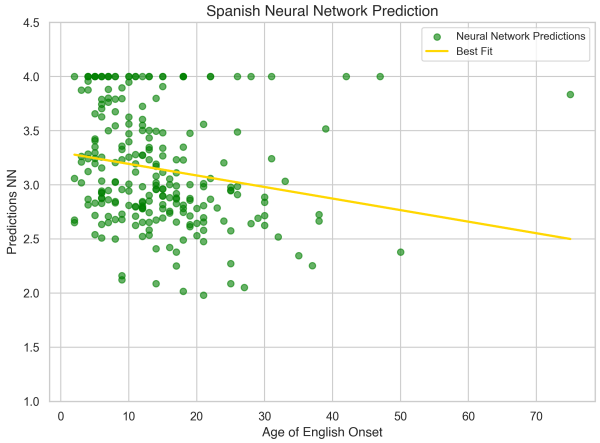

Figure 5.3: Spanish L2 Fluency Predictions against Age of English Onset

Figure 5.1 is a histogram showing the frequency distribution of fluency predictions across the L2 Spanish speakers. While the model shows a high number of predictions of the value 4, it also shows a relatively normal distribution of scores, predicting scores as low as two. Since scores show distribution across speakers, we can look at Figures 5.2 and 5.3 for correlation across our chosen metrics.

Figure 5.2 plots the speaker's "Years Learning English" on the x-axis and the predicted scores on the y-axis. If the model predicted fluency based on the years of English spoken, one

would expect a positive relationship between the two variables. The correlation coefficient was -0.01, showing little to no correlation between the years the reader has spoken English and the predicted fluency by the model. In light of this result, "Years Speaking English" does not appear to be a suitable proxy for the predicted fluency of a speaker within this model.

Figure 5.3 plots the speaker's "Age of English Onset" on the x-axis and the predicted scores on the y-axis. If the model predicted fluency based on the age of English onset, one would expect a negative relationship between the two variables. The correlation coefficient was -0.322, showing a low negative correlation between the age of English onset and the model's predicted fluency score. Between the two variables, "Age of English Onset" would be a better proxy for the fluency scores.

5.3 Discussion

The most immediate and striking result from Experiment 3 is that the model trained on spontaneous speech predicts a distribution of values across the tested L2 English speakers. This result should be compared with the values in Experiment 2. When trained on the same data, the model predicted the native English speakers to be more fluent than the L2 speakers, which is consistent with our expectations. Since there are no fluency scores to compare the L2 predictions to, we cannot comment too deeply on the nature and spread of the distribution. All we can say is that such a distribution exists.

Since the model predicts a distribution, we can ask which of our available metrics is the most accurate predictor of the model's fluency prediction. The experiment demonstrates that the "Age of Onset" variable is a better predictor of model fluency. Surprisingly, this mirrors the effect of these variables among humans. Tests on bilingual Chinese speakers of English have shown that the "age of acquisition" of the second language greatly impacts spoken fluency. Meanwhile, the duration of English speaking does not have a comparable effect on the judged fluency of human L2 speakers. This result adds credence to the model's

results and suggests that the model is picking on aspects of fluency judgments that human judges share.

Although the predictions' correlation with onset age is relatively weak, it shows sufficient strength to use as a proxy. The result makes sense since individual fluency ratings rely on many factors, of which onset age is merely one. Additionally, since we will be comparing the coefficients with each other in the next experiment, the objective strength of the correlation is less of an issue.

6. Experiment 4: Getting To The Point

We have answered the main concerns associated with mixing the read and spontaneous datasets through the previous experiments

- We saw that a narrow conception of fluency allows us to construct a model that that captures spontaneous speech proficiency.
- We showed that a model trained on spontaneous speech can generalize to predicting read speech
- We showed that the age of English onset can serve as a proxy for the fluency of L2 Read speech.

By combining this information, we can build a model that answers the questions we initially set out to ask.

In practice, the training data in official models tend to contain a large proportion of speech from native English speakers and a lower number of non-native English speakers (Cardwell et al. (2022) 2022:). This paper investigated the effects of the native language of L2 English speaker's automated fluency prediction. There are two factors associated with a native language that might influence the "fairness "of a fluency prediction. First is the presence of other L2 speakers with the same native language in the training data. The second is the "similarity "of the language to English. Experiment 4 attempts to simulate the environment of an official prediction model and test the fluency scores of L2 speakers from various native languages. This experiment compares the "fairness "of the model's predictions across four native languages, Spanish, German, Korean, and Farsi, using the "Age of Onset" variable as a proxy for actual fluency.

6.1 Methods

Experiment 3 showed that a speaker's "Age of English Onset" negatively correlates with their predicted fluency scores within the SAA, corresponding with its effect on human judgment. If the age of a speaker's English onset had a perfectly inverse correlation with spoken fluency, then the perfect model's predictions would exactly reflect the age of onset. Although such a world exists only in gross idealization, there is a sense in which the correlation between the model's fluency prediction of a speaker and their age of onset represents the fairness of the judgment. No element of the experimental setup would suggest that the correlation coefficients between the Age of English Onset and predicted fluency should differ between native languages. Therefore, a stronger correlation for a particular language implies that the language is being predicted more fairly and in line with their actual level of fluency. Using this setup, we can sidestep the absence of actual fluency scores in the SAA and directly examine the difference in judgment fairness across native languages. The model is trained on the entire Rated L2 Speech Corpus and the native subset of the SAA. All members in the native subset are assigned a score of 4. Most of the training data consists of native speakers, with relatively few non-native speakers. This setup mirrors industrial models such as the SpeechRater and the DET(Zechner and Evanini (2019) 2019, Cardwell et al. (2022) 2022). Once again, the speech-based features are input to the Neural Network model.

The model, once trained, is tested on four different languages. Each language represents a class so that they cumulatively represent all possible languages. Table 6.1 displays the chosen languages and their classification. The Rated L2 Speech Corpus, among other speakers, contains 14 Korean and 3 Spanish speakers of English. Meanwhile, the corpus contains no speech from native German or Farsi speakers. Additionally, there is an intuitive sense that Spanish and German are more similar to English than Farsi and Korean. These four languages are candidates to represent all the other languages that fit into these categories. The testing data consists of all SAA data for each of the four languages in Table 6.1. The

	Similar to English	Not Similar to English
In Training Data	Spanish	Korean
Not in Training Data	German	Farsi

Table 6.1: Language Classification Based on Similarity to English and Presence in Training Data

correlation between predicted fluency and onset age will be compared across these languages.

6.2 Results

Figure 6.1 presents the results of Experiment 4. The dark colors represent languages that have representation within the training data, while the light colors represent languages absent. The languages considered similar to English are on the left, while those not considered similar to English are on the right. For instance, the bar for Farsi is light and on the right side since it is not present in the training data and is not similar to English.

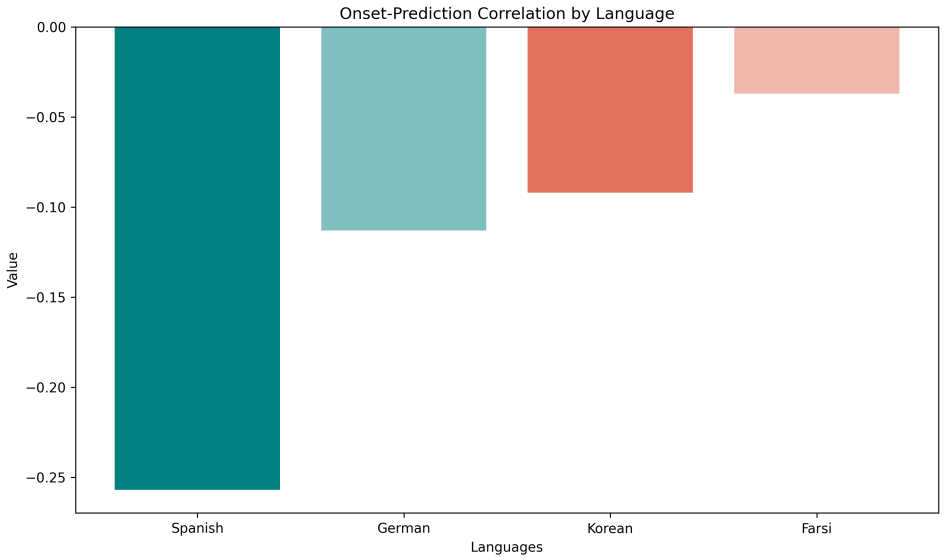

Figure 6.1: Language Onset-Fluency Correlation by Language

To explore the effects of training data representation, we should compare the performance of Spanish against German and Korean against Farsi. In both cases, the correlation is stronger in the languages with native speakers in the training data. Similarly, to find how

the similarity of the native language to English affects the predictions, we have to compare Spanish against Korean and German against Farsi. Therefore, the correlation is affected by the presence of the native language in the training data and its similarity to English.

To check the interaction effects between the two variables, a linear regression was conducted. A regression model was trained to predict the score using the speaker's onset age and an indicator of their native language's position. For both indices, the model was given a 1 if it satisfied the condition and a 0 if it did not. That way, a native speaker of Spanish had an index of [1,1] while a native speaker of Farsi had an index of [0,0]. Then, the coefficients of those two indices were checked to see the amount that it contributed to the fluency prediction. The score of the language similarity coefficient was 0.496 while the score of the dataset coefficient was 0.570.

6.3 Discussion

Experiment 4 tested the effect of two factors related to a speaker's native language on their L2 English fluency scoring. Both tested factors showed measurable effects on the correlation between the model's scores and the speaker's onset age, explaining the difference in coefficient values between Spanish — which has effects from both features — and Farsi, which has neither. From this limited experiment, German — similar to English but not in the dataset — had a higher coefficient value than Korean, which was in the dataset but not "close "to English. When testing for interaction effects, both factors had similar coefficients, but the presence of the language in the dataset seemed to have a higher effect on the fairness of the model than the similarity to English.

The effect of representation in training data is quite intuitive and rather unsurprising. When spoken over a large area or multiple discontinuous regions and cultures, a speaker's language acquisition is bound to be influenced by several regional factors. Considering its sheer number of native and non-native varieties, this is especially true of English. When a model predicts fluency scores, it extrapolates a conception of fluency based on the ratings

assigned to the training data. However, since there are different dialects of English, each dialect's conception of fluency varies slightly. Ideally, the model would have sufficient representation for all dialects of English. However, the presence of gold-standard ratings of an L2 English speaker of a particular native language within the training data increases the probability of a fair prediction since it has information about how fluency presents in that dialect. These results underscore the necessity of including a wide variety of English dialects from speakers of different native languages and regions.

Meanwhile, one must not conflate the effect of similarity to English with ease of language acquisition. This experiment does not say — and does not aim to say – anything about how easy it is for an L2 speaker to acquire English based on their native language. Instead, it explores how their perceived and judged fluency can differ through features of their native language that show much overlap with English. For instance, most of the model's training data consisted of native English speakers from the United Kingdom. When speaking English, native German speakers might sound more like these speakers than native Farsi speakers do. They might, therefore, receive a minimal increase in fairness, similar to the effect of the training data representation. One must, however, refrain from viewing this as over-scoring such speakers. After all, whether something can be "overly fair "is a question more suited to an undergraduate thesis in Philosophy.

7. Conclusion

This paper initially set out to ask questions about the effect of an L2 English speaker's native language on their automated fluency prediction. To simulate the architecture and models of official prediction, we wanted to use a combination of native and L2 speech to train the data and L2 speech to test the model. The two datasets available each had their issues that required reconciliation. Although the L2 Rated Speech Corpus had human-judged ratings, the dataset was too small and did not contain any native speakers. This feature is important to include in a fluency prediction model, since native speakers tend to be the gold standard for fluency within a langauge. Conversely, the Speech Accent Archive did not contain any fluency ratings. They also had respective modalities, with the former consisting of spontaneous and the latter of read speech. I had to conduct several checks on the data to answer the question reliably without any unwanted confounds interfering. Through Experiments 1-3, I ensured that a narrow description of fluency would be sufficient for both the Rated L2 Speech Corpus and the Speech Accent Archive and that the model could extract fluency from each dataset. Experiment 4 tested the original question. We also showed how the age of a speaker's language acquisition can be a proxy for fluency. The similarity of the native language to English and whether the language was present in the training data improved the correlation between age of onset and fluency. Finding this in the model was heartening, as it mirrored trends observed in human fluency judgments. While this does not conclusively prove that the model uses similar fluency methods as humans, it supports the idea.

The experiment did not set out to make a judgment about the most effective models or features to use for fluency judgments. The regression models performed better in a low-data context than the Neural Network. However, the neural model was capable of much better

generalization from across spontaneous to read speech. Due to this, we choose the neural model for the remaining experiments. However, this does not necessarily mean that neural networks are better for fluency prediction in all cases than the regression models.

It is essential to acknowledge the limitations of this study, many of which were due to data scarcity. An ideal replication of this experiment would involve a substantial corpus of spontaneous speech data, collected from a diverse range of native and L2 speakers, all of whom would have gold-standard ratings by expert human judges. Additionally, a more formal and objective metric for measuring the similarity of a native language to English would be beneficial. With these enhancements, the experiment would be more robust, and we could arrive at a more definitive conclusion regarding the performance of different models.

Overall, this experiment has demonstrated two separate ways in which an L2 speaker's native language can affect their automated fluency prediction. Many of these biases are inherent to fluency judgment, whether by humans or algorithms. Therefore, this result should not reflect a negative attitude toward introducing automation in fluency prediction. Instead, it should highlight areas of concern to ensure equitable assessment across all non-native English speakers.

8. Appendix A

This appendix enumerates each feature used in the experiment based on their category.

Table 8.1: Summary of Features by Category

Category	**Features**
Transcription-Based Features	Flesch-Kincaid Readability Score, Rarity (Inverse Log Probability of Word), Average Syntactic Depth
Phonetic Features	Words Per Minute, Syllables Per Minute, Number of Pauses, Average Length of Continuous Speech (Speech without pauses), Voiced Segment Length (Mean, Standard Deviation), Unvoiced Segment Length (Mean, Standard Deviation)
acoustic Features	Jitter (Mean, Standard Deviation), Shimmer (Mean, Standard Deviation), Spectral Flux (Mean, Standard Deviation), Cepstral Coefficients (Mean, Standard Deviation for each of the four values)

9. Appendix B

This table represents the grading rubric for the TOEFL iBT. It had also been adopted for fluency grading within the L2 Rated Speech Corpus. In both Experiments 1 and 2, Native speakers are automatically assigned a score of 4. This is because the TOEFL iBT is meant to be taken by non-native speakers, who can be assumed to be less fluent than native speakers.

Score	**General Description**
0	Speaker makes no attempt to respond or response is unrelated to topic.
1	The response is very limited in content and/or coherence is only minimally connected to the task or speech is largely unintelligible.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech although problems with delivery and/or overall coherence occur, meaning may be obscured in places.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas.
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse.

Bibliography

Anderson-Hsieh, Janet and Kenneth Koehler. ???? The effect of foreign accent and speaking rate on native speaker comprehension 38(4). 561–613. ISBN: 0023-8333 Publisher: Wiley Online Library.

Beinhoff, Bettina. 2014. Perceiving intelligibility and accentedness in non-native speech: A look at proficiency levels. vol. 5, 2014. Issue: 2014.

Berg, Thomas. 1999. Stress variation in british and american english 18(2). 123–143. Publisher: Wiley Online Library.

Cardwell, Ramsey, Geoffrey T LaFlair and Burr Settles. 2022. Duolingo english test: Technical manual .

Chan, May Pik Yu, June Choe, Aini Li, Yiran Chen, Xin Gao and Nicole R Holliday. 2022. Training and typological bias in asr performance for world englishes. In *Interspeech*, 1273–1277.

Cho, Hyesun. 2016. Variation in vowel duration depending on voicing in american, british, and new zealand english 8(3). 11–20.

Eberhard, David M., Gary F. Simons and Charles D. Fennig. 2023. *Ethnologue: Languages of the world*, vol. Twenty-sixth edition. SIL International. URL `http://www.ethnologue.com`.

Housen, Alex, Folkert Kuiken and Ineke Vedder. 2012. Complexity, accuracy and fluency 32. 1–20. Publisher: John Benjamins Publishing.

Kim, Hyunah, Liam Hannah and Eunice Eunhee Jang. ???? Using acoustic features to predict oral reading fluency of students with diverse language backgrounds. In *Collated*

papers for the ALTE 7th international conference, madrid, vol. 198.

Laan, Gitta PM. 1992. Perceptual differences between spontaneous and read aloud speech. In *Proc. of the institute of phonetic sciences amsterdam*, vol. 16, 65–79.

Langenkamp, Max, Allan Costa and Chris Cheung. 2020. Hiring fairly in the age of algorithms. *arXiv preprint arXiv:2004.07132* .

Lennon, Paul. 1990. Investigating fluency in EFL: A quantitative approach 40(3). 387–417. Publisher: Wiley Online Library.

Loukina, Anastassia, Klaus Zechner, Su-Youn Yoon, Mo Zhang, Jidong Tao, Xinhao Wang, Chong Min Lee and Matthew Mulholland. 2017. Performance of automated speech scoring on different low-to medium-entropy item types for low-proficiency english learners 2017(1). 1–17. Publisher: Wiley Online Library.

Nakamura, Masanobu, Koji Iwano and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance 22(2). 171–184. ISBN: 0885-2308 Publisher: Elsevier.

Ngueajio, Mikel K and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International conference on human-computer interaction*, 421–440. Springer.

O'Donnell, Renata M. 2019. Challenging racist predictive policing algorithms under the equal protection clause. *NYUL Rev.* 94. 544.

Park, Yena, Ramsey Cardwell, Sarah Goodwin, Ben Naismith, Geoffrey T LaFlair, Kai-Ling Lo and Kevin P Yancey. 2023. Assessing speaking on the duolingo english test.

Pinget, Anne-France, Hans Rutger Bosker, Hugo Quené and Nivja H De Jong. 2014. Native speakers' perceptions of fluency and accent in l2 speech 31(3). 349–365. Publisher: Sage Publications Sage UK: London, England.

Weinberger, Steven H and Stephen A Kunath. 2011. The speech accent archive: towards a typology of english accents. In *Corpus-based studies in language use, language learning, and language documentation*, 265–281. Brill.

Yoon, Su-Youn, Lisa Pierce, Amanda Huensch, Eric Juul, Samantha Perkins, Richard Sproat and Mark Hasegawa-Johnson. 2009. Construction of a rated speech corpus of l2 learners' spontaneous speech 26(3). 662–673. Publisher: JSTOR.

Zechner, Klaus and Keelan Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.