L2 English Learners Automatic Classification Methods *

Stella Podgornik[a]

Department of Linguistics, University of Washington
8210 Padelford Hall A210 Seattle, WA 98195, United States
stellap@uw.edu

Summary. This study analyses different features that would enable classifiers to detect language levels in adult second language (L2) English Learners. 46 different speech samples from users speaking 15 different L1 native languages were selected from the Learning Prosody in a Foreign Language (LeaP) corpus, Gut (2004). Using different groupings of features from the spoken L2 secondary language (English), a Support Vector Machine (SVM), was trained and the speakers were classified into three different categories: c1, c2, and s1. The categories used correspond to beginner, intermediate, and advanced levels of the target secondary language, English. These are the same categories used by the human annotators, and were also used for ease of comparison and consistency.

The features are grouped into four different sub–categories: sentence, syllable, duration, and pitch. Surprisingly, most of the pitch features had no effect on the accuracy. A small common word list was also used, please see Table 13. The edit distance measure of the sentences with the common words removed had a positive effect; measurable differences could be found with and without the common words included in the sentences. Due to the small size of the training and testing sets, it was found that the different groupings of the L1 languages of the speakers had a significant effect on the accuracy of the classification

* This research project would not have been possible without the support of many people. The author wishes to express her deepest gratitude to her supervisor, Prof. Gina–Anne Levow, who was abundantly helpful and offered invaluable assistance, support and guidance. Gratitude is due to Dr. Michael Tjalve, whose knowledge and assistance helped this study succeed. A special thank you also goes to my fellow student, Leanne Rolston, former technical writer and editor.

University of Washington Working Papers in Linguistics (UWWPL). Volume 30 (2013).

predictions. Certain combinations of L1 training and test sets had a higher accuracy depending on the L1 languages groupings. The classification predictions had a variance as much as 40%.

Keywords: Language Learning, Second Language Acquisition, Automatic Classification in SLA

1 Background and Introduction

1.1 Background

This study analyses different features that would enable classifiers to detect language levels in adult second language (L2) English Learners. 45 different speech samples from users speaking 14 different L1 native languages were selected from the Learning Prosody in a Foreign Language (LeaP) corpus, Gut (2004). Using different groupings of features from the spoken L2 secondary language (English), a Support Vector Machine (SVM), was trained and the speakers were classified into three different categories: c1, c2, and s1. The categories used correspond to beginner, intermediate, and advanced levels of the target secondary language, English. The native (L1) English speakers were not classified into any of the classification groups.

The categories are similar to the ones used by the human annotators. This method was utilized for ease of comparison and consistency between the system from this project and the human annotators. This system attempted to reproduce as accurately as possibly the same classification labels to the L2 speakers as given by the human annotators. The human annotators were used as the baseline for evaluation purposes.

1.1.1 Second Language Learning Requirements Children will gradually adjust their speech production as they mature, approximating a combination of the speech patterns produced by their parents and peers. This is how most people learn their first or native language, their L1 language, Gleason (1993). Learning a L1 language is done very gradually, and is usually incorporated as part of the child development process. Learning a second language is usually done after childhood. In order to learn a second language (an L2 language) there are three requirements, and without them a person will experience a difficulty or even an inability in learning a second language, Wong-Fillmore (1991):

- a motivation and a need to learn a second language;
- access to native speakers of the target language who are willing to give feedback, comment, and listen to repetitions ; and,
- a social setting where the L2 learner may practice on a frequent and persistent basis in the target L2 language.

1.1.2 Second Language Learning Obstacles Many L2 English learners have very little access to the second and third requirements, Wong-Fillmore (1991). According to many different studies on second language acquisition, the length of time exposed to and using the target language has a very direct effect on proficiency in the target language. The amount of practice in the L2 target language helps tremendously. Unfortunately, many L2 English learners learn English via Computer Aided Language Learning (CALL) software or in class room settings that are not in areas heavily populated by native English speakers. They hear the English words spoken incorrectly, and then repeat those same mistakes.

There are many differences in phoneme, duration, and even tone in the world's languages. The differences between the different native L1 languages can make it hard to find similarities between the speakers making different productions of English. This study has proved similarities can be found, quantified, and used to compare the similarities and differences in different levels of L2 English learners.

1.2 LeaP Corpus

The corpus being used is the LeaP corpus, collected in the LeaP project at the University of Bielefeld, Germany, from May 2001 to July 2003, Gut (2004). The LeaP project acquired data from non-native speakers ages 21–60 of English and German with a concern towards the acquisition of prosody. All recordings were digitised on a computer with 48kHz and 16 bit. Manual and automatic methods were used to annotate the data on 8 different tiers: pitch, tones, segments, syllables, words, phrasing, parts-of speech and lemmata. There are 359 annotated file with 131 different speakers speaking 32 different native languages. There are 18 recordings with native speakers.

2 Previous Work

Previous work most closely related to detecting L2 proficiency levels in English has been in a few different areas areas, and not all of it could be utilized for this project. In the pure linguistic field of second language acquisition, Flege et al (1997) looked at the effects of different L1 languages had on the vowel perception and production of English. The aim of the Flege's study was to better understand how experience can have an affect on adults' production of L2 English vowels. They found the L2 English speakers ability to accurately produce and perceive the English vowels improved as they gained in experience speaking English,

but their accuracy not only depended on experience but also on their L1 background. Flege's study used 20 speakers of 4 L1 different languages using a manual analysis method. In comparison to Flege's study, this project is looking at a larger sample size of study of over 40 participants with 11 different L1 languages using an automatic classification method.

Using a computational linguistics approach with various features: Maximum likelihood linear regression (MLLR), phone N-gram, prosodic, and word N-gram, Shriberg et al (2008) were able to detect whether a person is an L1 or L2 speaker of English in different SVM classifiers. This particular study used data and suitable models available for American English from prior work in speech and speaker recognition. This thesis paper and subsequent study had no known English models, but the LeaP corpus (Gut (2004)) did have native English data with which to compare the non–native English data to. Shriberg et al (2008) proved it was possible to automatically analyse a person's speech patterns and determine their fluency as a L2 speaker of English with a decent accuracy rate. There was a 16% performance gain in equal error rate (EER) and a 23% performance gain in the detection cost function (DCF) from the National Institute of Standards and Technology (NIST) . EER indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the equal error rate value, the higher the accuracy of the system.

The work of Chen et al (2010) is the closest and the most relevant to this study. They used outputs from a classifier using vowel space characteristics: f1–f2 distance, vowel area, and vowel dispersion, and compared them to human annotation scores of the speakers English proficiency. Chen et al (2010) were able to show how the acoustic analysis of vowel space for L2 English speakers could be used in an automated analysis, and how even the most highly proficient L2 English speakers had significant differences from the L1 speakers of English. In this thesis study, the final ranking of the non–native speech will also have a basis in the human annotation scores. While Chen et al (2010) used vowel features in order to score their users, vowel features was not used in this thesis study in order to find the native and non–native speaker differences. This thesis study looks at other acoustical features which also prove highly relevant in evaluating L2 English speakers in their proficiency of English.

D. Higgins and Williamson (2011) explain their work on *SpeechRater*^{SM}, an automatic system which scores L2 English speak-

ers in their proficiency of English. Their study built on previous work in the automated scoring of test responses for the Educational Testing Service (ETS) using highly unpredictable responses in their particular study. *SpeechRater*^{SM} uses a three-stage process, with the first stage using a filtering model to detect and only use responses which it deemed usable and able to assign a score to. D. Higgins and Williamson (2011) worked on the filtering model and proved it was feasible to develop a filtering model which was able to distinguish responses which would get good scores (1–4) from zero scores or "technical difficulty" ratings. D. Higgins and Williamson (2011) proved that using a pre-processing filter to separate "good" data (items with scores from 1–4) from "bad" data (items with zero scores) gave them decent accuracy scores which agreed with the human scores. This thesis study will not look at free speech, but instead will look at read speech data in order to determine the L2 speaker's proficiency in English. Free speech is recorded spontaneous speech from the speaker, and read speech is recorded speech from a speaker reading from a set script. Similar to D. Higgins and Williamson (2011), this thesis study will use a filtering method in the form of a common word list to differentiate a set common words, Table 13, from other words in the English language which might not be as relevant to look at.

3 Procedures and Measurements

Speech recognition techniques are used by most computerized learning systems to access the pronunciation quality for language learners word pairs or utterances. They compare the recognized spoken words with known standards of pronunciation. This project will try to detect the possible pronunciation errors in the phoneme level between the native (L1) and second language (L2) speakers, and will attempt to classify the L2 speakers on their levels of English proficiency. Disfluencies, pitch, spoken sentence accuracy, edit distance measurements of the L2 English learners speaking English will be compared with known native English speakers for cross–language phonological comparisons. Word and syllable durations will also be used as features in the cross–language comparisons. The L2 English speakers came from approximately 11 different countries and speak 11 different L1 languages other than their L2 English.

3.1 System Architecture

There were a few different steps for the overall system. The first step required writing a script in Praat, a vocal analysis program. Several modules were written in Java in order to analyze the text files produced from Praat. There were two different variations of analysing the data: one approach only analysed one test configuration and the other method analysed 50 randomly produced configurations. The one testing configuration was used in order to have a consistent, easily verifiable configuration for tests. The randomly produced configurations were used in order to find the accuracy dispersion between minimum and maximum accuracies resulting from the multiple different random configurations. The randomly produced configurations also helped determine the testing configuration, which was selected due to its accuracy result being close to the mean of the 50 different configuration accuracies. The same architecture was used for both an "automatic" and a "manual" configuration with SVMs used to classify the data into their different labels. In the field of machine learning, SVMs are supervised learning algorithms (a supervised algorithm compares calculated results to known data) which are used to analyse data and recognize patterns especially in classification and regression analysis. Due to its ease of use and past success, SVMs are believed to be the best classifier for this particular classification task.

3.2 Data Features Selected

The wave and TextGrid files that were selected all had syllable and tone tiers. All the files had a words tier of transcribed speech, and all the files with a syllable and tone tier had a segments tier as well. In Figure 1, the example shows a few blank spots on the segments tier. The blank sections in the segments tier are annotated with a C or a V, but the Praat windows are not zoomed in at a small enough level to see the annotation because the sections are very small. Not all the files had all the tiers. Some of the TextGrid files only had one annotation tier, some had several tiers, and some files had all the tiers displayed in Figure 1.

The reasons for the selections was the need for the formant, intensity and pitch information associated with the vowels, and retrieving this information from the files is very difficult without the vowel and consonant annotation. All these measurements are only taken in the vowel part of the syllable. Not all the languages or speakers described in the LeaP corpus, Gut (2004), were used for this study due to these

L2 Language Speakers		
L1 Language	Classes	Total
Arabic	c1	1
Chinese	c1	1
English	na	3
French	c1	1
German	c1 c2 s1 ot	11 4 7 1
Hungarian	c1 c2	2 2
Ibibio	ot	1
Igbo	ot	1
Korean	c1 ot	2 1
Persian	c1	1
Russian	c1	3
Spanish	ot	2
Yoruba	ot	1
Total		45
Used for L2 calculations		35

Table 1: L2 Language Speakers. There were 3 possible classes: c1, c2, s1. The na class was only used for training and combined with the s1 training classes. The na class was not classified for testing. The ot classes were not used in training, but only in testing for further evaluation at another time.

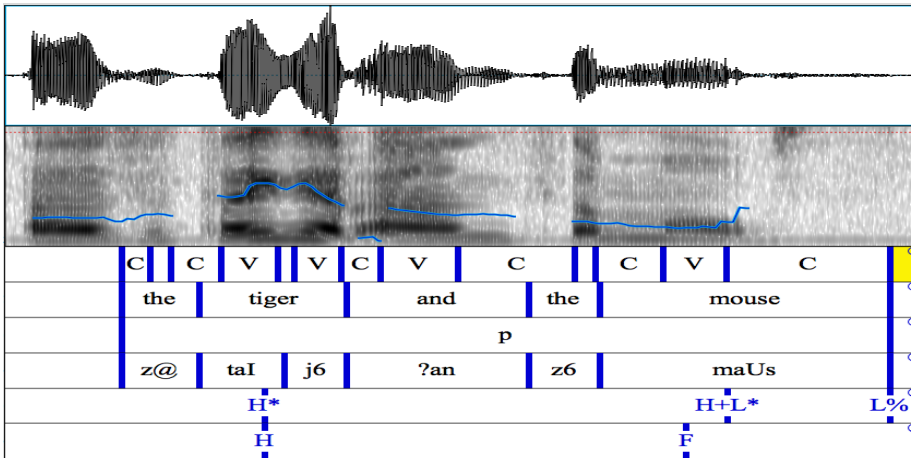

Figure 1: Example Annotated Wave File in Praat. L1 Arabic speaker.

constrictions. There were single word tables and free speech also available in the LeaP Corpus, but not used. Free speech is speech without any transcript or pre–meditation. The data used for this study is read data, which is read speech recorded from a speaker who is reading from a given transcript. This transcript is the same for all the speakers in corpus. Each speaker being evaluated is reading from the same exact script, and comparisons are calculated when the speakers deviate from that script. Table 1 describes the files used in the final evaluation.

Figure 1 is an example from an L1 Arabic speaker.

3.2.1 Edit Distance Formulas Edit Distance Formulas are also known as approximate string matching or fuzzy string searching and they are used to numerically calculate how to transform one string to another string or how close two strings are to each other. If the strings are either equal or both empty, the resulting distance measure with Damerau–Levenshtein will always equal 0, while the Cosine Similarity algorithm will always equal 1.

Damerau–Levenshtein The Damerau–Levenshtein distance formula is a set of elementary operations of one-character deletions, insertions, substitutions, and transpositions of adjacent characters. It is a non-weighted edit distance, where all operations have the same weight. The lower the result, the closer each string is to each other.

Cosine Similarity The cosine similarity method is a little more complicated in that it measures the similarity between two vectors by measuring the cosine of the angle between those vectors. The resulting distance is 1 when the angle is 0, and it grows smaller as the angle gets larger. The cosine calculation of the angle between two vectors will therefore determine if two vectors are roughly pointing in the same direction. Each string is transformed into a set of numerical objects or vectors, with each word given a numerical value on its importance in the overall document being analyzed. The more frequently a word appears, the higher its value. In the Cosine Similarity method, these resulting vector angles are compared to see how far apart each string is to another. The Cosine Similarity algorithm was edited to produce a calculation to be comparable to Damerau–Levenshtein, with both algorithms resulting in zero for identical strings.

3.2.2 Sentence Features Table 2 shows the Damerau–Levenshtein and Cosine Similarities calculations for the sample utterance: the spoken sentence, the sentence text from the story, and both edit distance measures.

spoken	a tiger and a a mouse were walking in a field when they saw a big lump of cheese lying on the ground
text	a tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground
scores	damlev 2.0, cossim 0.0643

Table 2: Sentence Features. Damerau–Levenshtein and Cosine Similarities calculations

3.2.3 Syllable Features In the syllables section, each rough phonemic transcription was compared to the comparable sentences from each of the native English speakers' phonemic transcriptions. For each speaker that read the following phrase "a tiger and a mouse were walking in a field", each passage from each speaker was individually compared to each L1 English speaker. There were three scores for each of the three L1 English speakers, and the score which had the nearest distance to one of the three L1 English speakers was kept for the overall L2 speaker feature vector. There were not enough native L1 English speakers to form a one "common person" amalgamation. As with the sentence

features, the Damerau–Levenshtein and Cosine Similarity algorithms were used for the phonemic transcription calculations.

The following sentence is an example of the phonemic syllabic transcription of the first few words of the story. The syllables are transcribed in SAMPA. SAMPA stands for Speech Assessment Methods Phonetic Alphabet, and it is a computer-readable phonetic script using 7-bit ASCII characters using only the keys from a standard keyboard.

```
a tiger and a a mouse were walking in a field
a, taI jer, ant, a, a, maUs, wer, wO kIN, In, ?{, fiKd, pause
```

3.2.4 Duration Features Both the duration and pitch features were normalized using the the zero normalization method first mentioned by Li and Porter (1988). Variations in duration and pitch can be removed by making the scores relative to the mean and variance. $L(x_i|S)$ is a score for a given speaker S and the given feature x_i where an overall utterance is denoted by X=$\{x_i\}$ $i \in$ [1,N]. $L(x|S)$ is the raw score, while the normalized score is $L_{norm}(x_i|S)$, Li and Porter (1988).

3.2.5 Pitch Features Although pitch measurements were computed at different points in the vowels by Praat, the pitch measurements used for the feature vector were all taken at the 50% point for the vowel. The pitch for each syllable, and the pitch statistics were taken into account for the final vector for each speaker dataset. During the testing phase of the experiment, individual pitch and sentence pitch features were disregarded in favor of the read passage pitch statistics. The same z-norm normalization procedure was used for both pitch and duration features.

3.3 Selecting and obtaining the data set

Praat, Boersma and Weenink (2011), was used to extract the numerical data from the wav and TextGrid files. A Praat script was developed to extract features from each wav and associated TextGrid file, and to output the information into a text file. Each language was separated into its own folder, and each language sample had its own file in the associated language folder. The Praat script stepped through the syllable tier in the TextGrid and picked out the following information associated with each syllable unit: word label of the syllable, syllable label, start and end times, tone if it was available, pitch minimum, pitch median, pitch maximum, and the f0, f1, f2 and f3 from the onset, midpoint and glide of the component vowel. To establish a baseline, only the read speech was processed and analyzed. The free form speech samples will be analyzed in a later project.

Automatic Feature Selection – Relief When looking at a large selection of features, some features may confuse and contradict each other. An individual feature may be useless by itself but really stand out and provide good data separation when combined with another feature. The Relief method is a multi–class example of a multivariate filter, and feature ranker. It ranks subsets of features versus ranking them individually. It can reduce noise and reduce feature redundancy in order to find the optimal set of features for the highest accuracy rates, Guyon and Elisseeff (2006). In the automatic configuration, the features were rearranged into the configuration Relief determined was the optimal configuration for data separation. The features in Table 3.3 are the top 10 features important to Relief.

The Rank column describes the features rank in the manually ordered feature vector, and the Feature No. column is the feature number in the full feature vector found in Appendix A. The numbers the Feature No. column in Table 3.3 correspond to the same feature number and same feature in Table 3.3. A copy of this table can also be found in Appendix A in order to easily compare the manual and the relief ordered full feature vectors.

Top 10 Relief (Automatic) Ordered Feature Vector		
Rank	Feature No.	Feature
1–2	3–4	Damerau–Levenshtein distances between sentences 3-4
3	156	Maximum word duration in sentence 7
4	133	Minimum word duration in sentence 5
5	144	Maximum word duration in sentence 6
6	123	Mediam word duration in sentence 4
7	11	Cosine Similarity distance in sentence 4
8	85	Minimum word duration in sentence 1
9	86	Mean word duration in sentence 1
10	47	Damerau–Levenshtein distance in syllablic transcription in sentence 1

Table 3: Top 10 Relief (Automatic) Ordered Features. A full breakdown of the feature vector and the Relief order feature vector can be found in Appendix A. The feature number in the table corresponds to the feature number from the fully featured SVM vector.

Manual Feature Selection The features selected manually were selected through trying many different configurations, a ”trial and error” approach. At first all the features were tried, and one by one, each feature array was added or discarded in trying to reach an optimal set of features with the highest accuracy rates. Unlike Relief, which gave outputs of the features individually but grouped the features in clusters which worked well together, the manual selection method gave the feature outputs in groups of similar features, i.e. the sentence features are together, the syllable features are together, etc... Each set of features were grouped with other sets of features which were similar. Table 3.3 has the top 29 of the features selected manually.

Manually Ordered Feature Vector		
Rank	Feature No.	Feature
1–7	27–33	Damerau–Levenshtein distances between sentences 1–7 without the common words
8–14	34-40	Cosine Similarity distance in sentence 1–7 without the common words
15	41	Ratio of Out of Vocabulary word to In Vocabulary words per passage

Table 4: Top 15 Manually Selected Feature Set. These features are not the full feature set due to the manual selection and elimination process. Some features in the automatic selection process were eliminated in the manual process. The feature number in the table corresponds to the feature number from the fully featured SVM vector.

4 Results and Analysis

In the Results chapter, the test configuration used in each of the automatic and manual sections refers to the experimental testing configuration. The same testing configuration dataset was used for both the automatic and manually configured tests. Table 2 illustrates both of evaluation methods together, and what each method was able to achieve at each combination of features. All features at any certain given point combines with all the features previous to it.

4.1 Automatic Configuration

4.1.1 Relief Ordered, Test Configuration In the Relief ordered, automatic configuration, all the features were used in an optimal configuration pattern found by Relief. Figure 3 shows the accuracy results

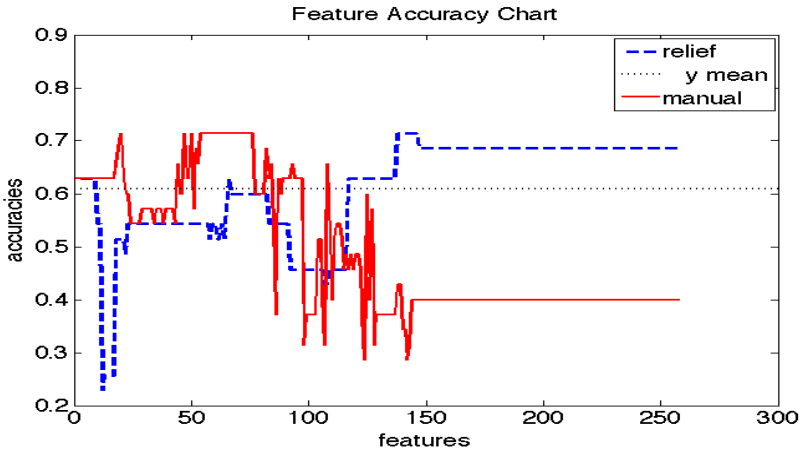

Figure 2: Automatic and Manual Accuracy Chart. The features on the x–axis correspond to the Rank column in Table A

that resulted from incrementally increasing through the Relief ranked and ordered features. The accuracy numbers rose and fell around the default class accuracy until it reached its highest value at 71%, at features 138–146. All the numbers in Figure 3 are for the test configuration.

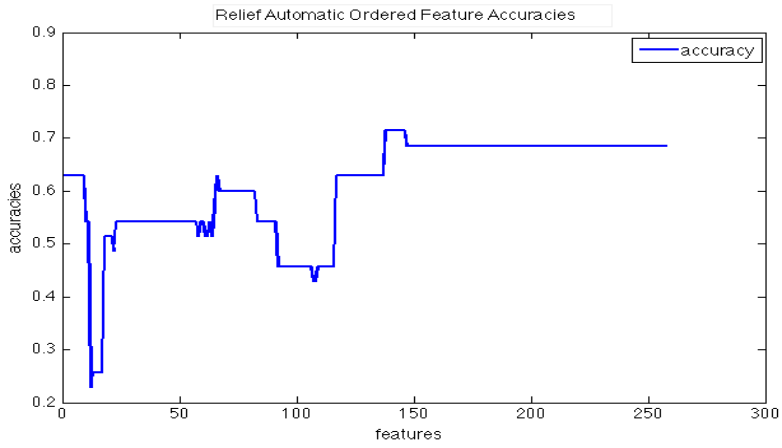

Figure 3: Relief Ordered Test Accuracy Chart. The features on the x–axis correspond to the Rank column in Table 3.3

4.1.2 Class Accuracies Both classes, c1 and s1, have positive relationships with the overall test accuracy. As each one set of accuracies go up, they all go up. Class c2 is not featured in the chart due to its constant value of "0". Classes c1 and s1 have a very minor negative correlation (the linear dependance between two variables X and Y) to each other. Relief ordered the features into an optimal configuration for class separation, and it found the optimal configuration for class separation between c1 and s1, while being unable to detect c2. Relief finds the optimal features for data separation, and will disregard the features which are not optimal. Due to the classifier being unable to detect c2, it is assumed the features needed to detect c2 are not optimal. The correlations between the classes and the overall test accuracy were computed using the Pearson product–moment correlation coefficient, also known as PPMCC or PCC, Rodgers and Nicewander (1988). The higher the score, the higher the correlation.

Accuracies	Correlations
overall vs c1	0.9178
overall vs s1	0.5263
c1 vs s1	-0.1456

Table 5: Relief Ordered – Class Correlations. Both c1 and s1 helped the overall testing configuration accuracy go up, as the overall accuracy improved, so did c1 and s1 individually. C1 and S1 had a very minor to almost no correlation to each other in this configuration.

4.2 Manual Configuration

4.2.1 Manually Ordered Test Configuration Figure 4 shows the accuracy results that resulted from incrementally increasing through the manually ordered features. For the first set of features 1–18, not enough information was available to the SVM to make a decision, so it defaulted to the result of the largest class, c1, about 62%. The accuracy numbers rose and fell around the default class accuracy until it reached its highest value at 71%, at feature 75. All the numbers in Figure 4 are for the test configuration. Unlike the Relief ordered and ranked features, the accuracy numbers for the manual configuration are spikier, and achieve its highest value with a smaller subset of features from the full feature vector. The manual configuration was able to achieve similar results to the Relief ordered configuration with less features. For easier comparison please see 2.

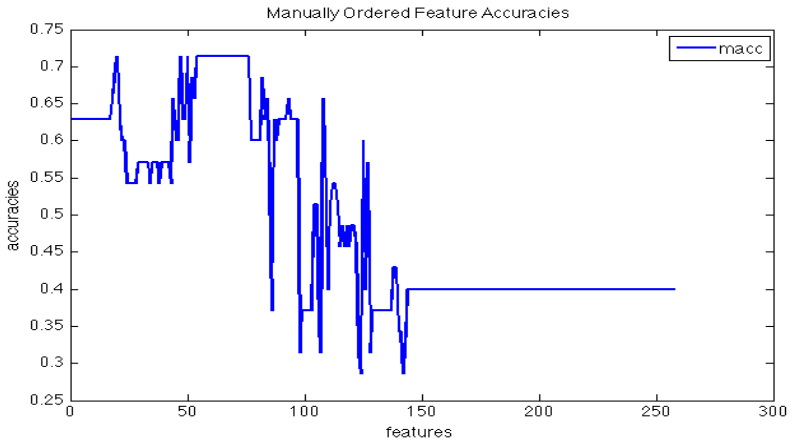

Figure 4: Manually Ordered Test Accuracy Chart. The features on the x–axis correspond to the Rank column in Table 3.3

4.2.2 Class Accuracies Unlike the Relief ordered features, the manually ordered and selected features were able to detect and separate out the c2 class from the other classes: c1 and s1. While c1 and c2 both had small positive correlations with the overall accuracy, s1 had a very slight negative correlation. Comparing the classes to each other, c1 vs c2, and c1 vs s1 had strong negative correlations, while c2 vs s1 had a good positive correlation.

Accuracies	Correlations	Accuracies	Correlations
overall vs c1	0.9142	c1 vs c2	–0.1166
overall vs c2	0.2160	c1 vs s1	–0.3416
overall vs s1	0.0418	c2 vs s1	0.5296

Table 6: Manually Ordered – Class Correlations. The overall accuracy had a strong correlation to c1‘s accuracy, and no correlation with s1. S1 improved its score as c2 improved, and c1 decreased.

In Table 6 it is obvious there is a strong correlation between the overall system accuracy and the accuracy for the c1 class. This does show the effectiveness of the manually selected features in being able to find a good configuration for data separation with less features. The surprising result of Table 6 is while the c1 class had minor negative correlations

to both the c2 and s1 classes, the c2 and s1 classes had a good positive correlation to each other.

4.3 Test Configuration Accuracies

The most important features in getting the highest classification accuracies for both the automatic and the manual configurations, and both their overall and class scores were the durational features. While in the overall, c1, and c2 accuracies, all the words' durational features, c2 in particular heavily relied on the common word durations in order to make its decisions. The overall, c1, and c2 also found the Damerau-Levenshtein distances between syllabic transcriptions and the best scored native English speakers important. The syllabic and the common word durations were particularly important in getting high accuracies in the s1 class. In Table 7 the dash between the numbers is trying to associate the rank of the feature with its respective configuration number in the fully featured SVM vector, Table A.

5 Highest Accuracy Features				
	Automatic		Manual	
No.	A. Accuracy	Rank – Feature	M. Accuracy	Rank – Feature
1	0.7142	138 – 52	0.7142	20 – 50
2	0.7142	139 – 143	0.7142	53 – 131
3	0.7142	140 – 124	0.7142	54 – 150
4	0.7142	141 – 76	0.7142	55 – 151
5	0.7142	142 – 101	0.7142	56 – 152

Table 7: Testing Accuracies – Overall System – Top 10. The Rank number refers to the feature's rank within either the automatic or the manual configuration. The feature numbers can be associated with their feature names in Appendix A. Before reaching the highest accuracy, for the highest first highest ranked feature, there were n-1 features in front of that specific feature in order to reach the highest accuracy numbers.

4.3.1 Class Accuracies In the class accuracies for the testing configuration, the automatic configuration in the same top 5 spots for both the overall system accuracies and the individual class accuracies. The manual configuration has different features in the top 5 spots for the overall system and the individual classes depending on the individual class.

The automatic configuration did not detect the c2 class, so all the information in Table 9 is the manual configuration. Due to the absence

C1 – 5 Highest Accuracy Features				
	Automatic		Manual	
No.	A. Accuracy	Rank – Feature	M. Accuracy	Rank – Feature
1	1.0	138 – 52	1.0	19 – 49
2	1.0	139 – 143	1.0	21 – 51
3	1.0	140 – 124	0.9545	107 – 238
4	1.0	141 – 76	0.9090	20 – 50
5	1.0	142 – 101	0.8636	18 – 48

Table 8: C1 – Testing Accuracies – Top 10. The Rank number refers to the feature's rank within either the automatic or the manual configuration. The feature numbers can be associated with their feature names in Appendix A. Before reaching the highest accuracy, for the highest first highest ranked feature, there were n-1 features in front of that specific feature in order to reach the highest accuracy numbers.

of the automatic configuration results, Table 9 has a slightly different format and shows the results for all the class sorted in descending order of the c2 class. The common word median in sentence 2 was the feature giving the best separation for c2. The common words in sentence 2 are "the, said, let me have it". The common syllable count per passage, the word and syllabic durations, and the Damerau-Levenshtein distances between syllabic transcriptions and the best scored native English speakers also proved important.

C2 – 5 Highest Accuracy Features Manual Configuration					
No.	Rank	Feature	C1	C2	S1
1	105	105	0.1818	0.8333	0.4285
2	46	46	0.7272	0.6666	0.7142
3	85	197	0.0909	0.6666	1.0
4	122	248	0.1363	0.6666	0.5714
5	49	127	0.7727	0.5	0.7142

Table 9: C2 – Testing Accuracies – Top 10. The automatic configuration did not differentiate the c2 class. Only the manual configuration is included in this table. The feature numbers can be associated with their feature names in Appendix A. Before reaching the highest accuracy, for the highest first highest ranked feature, there were n-1 features in front of that specific feature in order to reach the highest accuracy numbers.

S1 – 5 Highest Accuracy Features				
	Automatic		Manual	
No.	A. Accuracy	Rank – Feature	M. Accuracy	Rank – Feature
1	0.7142	138 – 52	1.0	85 – 197
2	0.7142	139 – 143	0.85714	108 – 239
3	0.7142	140 – 124	0.85714	109 – 240
4	0.7142	141 – 76	0.7142	46 – 118
5	0.7142	142 – 101	0.7142	49 – 127

Table 10: S1 – Testing Accuracies – Top 10. The Rank number refers to the feature's rank within either the automatic or the manual configuration. The feature numbers can be associated with their feature names in Appendix A. Before reaching the highest accuracy, for the highest first highest ranked feature, there were n-1 features in front of that specific feature in order to reach the highest accuracy numbers.

4.3.2 Language Accuracies In Table 11, each individual L1 language correlation is calculated to its matching class, to see how dependent the class accuracy is to that particular language's same class. The overall correlation shows how dependent the overall accuracy is to that particular languages class. Finally, the class and overall dependancies are shown at the end of each class to measure how related each class is to the overall accuracy.

5 Conclusion

In this study, the type of features with the highest overall impact on the system for both the automatic and manual configurations are the durational features. For the overall, c1, and c2 accuracies, all the words' durational features were important for the classifier to make its decision, while c2 heavily relied on the common word durations. Both the syllabic and the common word durations were particularly important in getting high accuracies for the s1 class. The c2 classified speakers were only detected by the manual configuration, and had zero results in the automatic configuration. Due to the small size of the data set, the testing and training configuration used had a very large effect on the overall accuracy of the system. Depending on which of the 50 random configuration selected for training and testing, there was as up to 40% fluctuation in the minimum and maximum testing accuracies. Due to this fluctuation, a training/testing configuration was selected close to the mean accuracy of the random configurations.

Language / Class Correlations					
L1 Language	Class	Class Correlation		Overall Correlation	
		Automatic	Manual	Automatic	Manual
arabic	C1	0.7725	0.8634	0.7333	0.9095
chinese	C1	0.7094	0.2283	0.8173	0.1669
french	C1	0.7268	0.7616	0.5656	0.7459
german	C1	0.8252	0.7542	0.8253	0.5379
hungarian	C1	0.3210	0.2090	0.1417	0.0915
korean	C1	0.8626	0.7432	0.7621	0.7324
persian	C1	0.6815	0.7353	0.6818	0.8113
russian	C1	0.8766	0.9095	0.8767	0.9359
C1	C1			0.9178	0.9142
german	C2	0.0	0.8830	0.0	0.2966
hungarian	C2	0.0	0.6052	0.0	–0.0489
C2	C2			0.0	0.2160
german	S1	1.0	1.0	0.5263	0.0418
S1	S1			0.5263	0.0418

Table 11: Language / Class Correlations. All these numbers were generated from the one testing configuration.

The system was able to classify speakers into the three different classes with good success due to the relatively small amount of features. In the work of Chen, Evanini and Sun formant and vowel features boosted the accuracy of their systems, Chen et al (2010), and I believe the addition of formant and vowel features could possibly boost this overall system accuracy above 80%. I would like to see the addition of phonetic features used in an overall automatic tutoring system or into CALL systems, where the system will be able to detect a speakers English ability, and then automatically tailor its teaching style to that particular speaker. Most L2 learners of English do not have ready access to a social support system of native English speakers who will be able to reinforce, and repeat the words with a correct pronunciation. Most L2 English learners have the most contact with other L2 English learners who are making the same pronunciation mistakes they are. Similarities can be quantified and used to automatically detect their proficiency in English.

References

Boersma P, Weenink D (2011) Praat: A system for doing phonetics by computer

Chen L, Evanini K, Sun X (2010) Assessment of non-native speech using vowel space characteristics. IEEE Workshop on Spoken Language Technology

D Higgins KZ X Xi, Williamson D (2011) A three-stage approach to the automated scoring of spontaneous spoken responses. Computer Speech and Language 25(1):282–306

Flege J, Bohn O, Jang S (1997) Effects of experience on non-native speakers production and perception of english vowels. Journal of Phonetics 25:437–470

Gleason J (1993) The Development of Language. Macmillan Publishing Company, New York

Gut U (2004) The leap corpus. URL `http://www.phonetik.unifrieburg.de/leap/LeapCorpus.pdf`

Guyon I, Elisseeff A (2006) An introduction to feature extraction. In: Feature Extraction, Foundations and Applications, Springer, pp 1–12

Li K, Porter J (1988) Normalizations and selection of speech segments for speaker recognition scoring. Proceedings of ICASSP 88 1:595–598

Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. The American Statistician 42(1):59–66

Shriberg E, Ferrer L, Kajarekar S, Scheffer N, Stolke A, Akbacak M (2008) Detecting nonnative speech using speaker recognition approaches. Proc Odyssey Speaker and Language Recognition Workshop

Wong-Fillmore L (1991) When learning a second language means losing the first. Early Childhood Research Quarterly 6:323–346

A Full SVM Feature Vector

Full SVM Feature Vector	
Sentence Features	
1–7	Sentence: Damerau–Levenshtein Distances Sentences 1-7 from the original text
8–14	Sentence: Cosine Similarity Distances Sentences 1-7 from the original text
15–20	Sentence: Damerau–Levenshtein Distance Statistics (in order): Summation, Minimum, Maximum, Median, Mean, Standard Deviation
21–26	Sentence: CosineSimilarity Distance Statistics (in order): Summation, Minimum, Maximum, Median, Mean, Standard Deviation
27–33	Sentence: Damerau–Levenshtein Distance Non–Common Word Sentences 1-7 from the original text
34–40	Sentence: Cosine Similarity Distance Non–Common Word Sentences 1-7 from the original text
41	Passage: Ratio of the words said Out of Vocabulary to In Vocabulary (only the words from the story text are considered In Vocabulary)
42	Passage: Ratio of the counted words In Vocabulary to the total amount of the characters for the words
43–45	Passage: Count of the Out of Vocabulary words, the count of the In Vocabulary words, the count of the articles per each sentence
Syllable Features	
46	Passage: Common syllable count
47–53	Sentence: Damerau–Levenshtein Distances syllalble transcribed sentences 1-7 from the original text
54–60	Sentence: Cosine Similarity Distances syllalble transcribed sentences 1-7 from the original text
61–67	Sentence: Damerau–Levenshtein Distance syllable transcribed statistics (in order): Summation, Minimum, Maximum, Median, Mean, Standard Deviation
68–74	Sentence: CosineSimilarity Distance syllable transcribed tatistics (in order): Summation, Minimum, Maximum, Median, Mean, Standard Deviation
75	Passage: Count of all the syllables per sentence which were In Vocabulary
76	Passage: Count of all the syllables per sentence which were Out of Vocabulary
Pitch Features	
77–82	Passage: pitch statistics (in order): Summation, Minimum, Maximum, Median, Mean, Standard Deviation
83	Passage: the total length of the pitch array
Duration Features	
84–89	Sentence 1: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
90–95	Sentence 1: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
96–101	Sentence 2: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
102–107	Sentence 2: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
108–113	Sentence 3: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
114–119	Sentence 3: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
120–125	Sentence 4: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation

Full SVM Feature Vector	
126–131	Sentence 4: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
132–137	Sentence 5: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
138–143	Sentence 5: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
144–149	Sentence 6: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
150–155	Sentence 6: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
156–161	Sentence 7: Word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
162–167	Sentence 7: Common word durations per sentence: Maximum, Minimum, Mean, Median, Sum, Standard Deviation
168–174	Sentences 1–7: Average common word count per sentence
175	Passage: Average common word duration for passage
176–180	Sentence 1: Syllable durations: last five per sentence
181–186	Sentence 1: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
187–191	Sentence 2: Syllable durations: last five per sentence
192–197	Sentence 2: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
198–202	Sentence 3: Syllable durations: last five per sentence
203–208	Sentence 3: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
209–213	Sentence 4: Syllable durations: last five per sentence
213–218	Sentence 4: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
219–223	Sentence 5: Syllable durations: last five per sentence
224–229	Sentence 5: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
230–234	Sentence 6: Syllable durations: last five per sentence
235–240	Sentence 6: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
241–245	Sentence 7: Syllable durations: last five per sentence
246–251	Sentence 7: Syllable duration statistics per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
252–257	Passage: Pause durations, Maximum, Minimum, Mean, Median, Sum, Standard Deviation
258	Passage: Common Word Count

Table 12: **Full SVM Feature Vector**. Row by row breakdown of each vector feature for the SVM.

B Common Words Used

a	able	about	across	after	all	almost
also	am	among	an	and	any	are
as	at	be	because	been	but	by
can	cannot	could	dear	did	do	does
either	else	ever	every	for	from	get
got	had	has	have	he	her	hers
him	his	how	however	i	if	in
into	is	it	its	just	least	let
like	likely	may	me	might	most	must
my	neither	no	nor	not	of	off
often	on	only	or	other	our	own
rather	said	say	says	she	should	since
so	some	than	that	the	their	them
then	there	these	they	this	tis	to
too	twas	us	wants	was	we	were
what	when	where	which	while	who	whom
why	will	with	would	yet	you	your

Table 13: **Common Word List**. The edit distance measure of the sentences with the common words removed had a positive effect; measurable differences could be found with and without the common words included in the sentences.

C Automatically Selected Features SVM Vector

Relief Automatically Ordered Feature Vector	
Ranks	Feature Numbers
1 – 20	3, 4, 156, 133, 144, 123, 11, 85, 86, 47, 8, 45, 126, 9, 142, 94, 10, 53, 82, 102
21 – 40	49, 121, 43, 44, 125, 59, 109, 106, 50, 26, 130, 118, 18, 97, 89, 57, 15, 77, 30, 155
41 – 60	116, 122, 21, 24, 105, 140, 40, 69, 110, 80, 127, 148, 38, 128, 149, 68, 58, 139, 66, 65
61 – 80	51, 88, 63, 60, 56, 36, 79, 99, 152, 91, 131, 135, 145, 33, 150, 20, 104, 1, 23, 2
81 – 100	35, 96, 70, 146, 75, 29, 17, 74, 64, 147, 22, 54, 153, 100, 132, 32, 6, 83, 117, 136
101 – 120	87, 92, 95, 31, 19, 41, 61, 34, 5, 67, 98, 14, 108, 93, 107, 13, 46, 90, 151, 37
121 – 140	28, 39, 12, 42, 134, 129, 48, 27, 154, 113, 7, 55, 73, 112, 84, 115, 16, 52, 143, 124
141 – 160	76, 101, 81, 72, 103, 120, 78, 137, 25, 141, 138, 71, 114, 111, 62, 119, 157, 158, 160, 164
161 – 179	43, 161, 162, 163, 165, 159, 166, 167, 168, 170, 171, 172, 173, 174, 175, 176, 177, 178, 169

Table 14: **Relief Ordered Feature Ranking**. All the missing numbers 179–258 from the full feature vector are placed at the end of the feature vector and had very little impact on the automatic vector accuracies.

D Manually Selected Features SVM Vector

Manually Ordered Feature Vector		
Rank	Feature No.	Feature
1–7	1–7	Damerau–Levenshtein distances between sentences 1–7 without the common words
8–14	8–14	Cosine Similarity distance in sentence 1–7 without the common words
15	41	Ratio of Out of Vocabulary word to In Vocabulary words per passage
16	46	A count of all the common word syllables per passage
17-23	47 – 53	Damerau–Levenshtein distances between syllabic transcriptions 1–7 and the best scored native English speaker
24 – 29	77 – 82	Pitch statistics for the words per passage: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
30	83	Numbers of pitches per passage
30–35 36–41 42–47 48–53 54–59 60–65	90–95 102–107 114–119 126–131 150–155 162–167	Word Durations per sentence: maximum, minimum, mean, and median Common Word durations per sentence: summation and standard deviation
66–72	168–174	Common word count per sentence
73	175	Average common word duration per passage
74–79 80–85 86–91 92–97 98–103 104–109 110–115	181–186 192–197 203–208 213–218 224–229 235–240 246–251	Duration statistics for the syllables per sentence: Summation, Minimum, Maximum, Median, Mean, Standard Deviation
116	252	Maximum pause duration per passage
117	252	Minimum pause duration per passage
118	256	Summation of all the pauses per passage
119	257	Standard deviation of all the pauses per passage
120	258	Common word count for the passage

Table 15: **Full Manually Selected Feature Set**. The missing features from the full feature vector are placed at the end of the vector and had very little impact on the manual vector accuracies.