Supervised Audio-Based Classification Framework for Fluency and Pronunciation Evaluation in Non-Native English Speech

Beifeng Wu*, Yongjin Hu
School of Foreign Studies, Suzhou University, Suzhou, Anhui, 234000, China
E-mail: wubeifeng2024@163.com
*Corresponding author

Keywords: image processing, deep residual adaptive networks, multimedia communication, graphical design automation, deep learning for image transformation

Received:

This paper presents a supervised machine learning framework for automated assessment of second-language (L2) English fluency and pronunciation, using audio recordings from Latin-American English learners. The framework extracts key acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), Zero Crossing Rate (ZCR), root mean square energy, and spectral features, from segmented audio samples. Multiple classification models — such as Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (kNN), and Convolutional Neural Networks (CNN), were trained to classify proficiency levels (basic, intermediate, advanced) and detect specific pronunciation errors. In addition, regression models, including Random Forest Regressor, were applied to predict continuous pronunciation quality scores. The study used a carefully curated dataset comprising over 18,000 audio segments, expanded through data augmentation techniques such as time shifting and playback speed variation. Experimental results show that the SVM classifier achieved over 94% accuracy in fluency classification, while the kNN model reached up to 99.9% accuracy in pronunciation evaluation. The Random Forest Regressor achieved a coefficient of determination (R^2) exceeding 0.92 for predicting continuous pronunciation scores, demonstrating the framework's robustness and scalability. These findings highlight the potential of data-driven, non-speech-recognition-based approaches for scalable, automated, and accurate L2 speech assessment.

Povzetek: Narejena je metoda za analizo naglasov in kvalitete govora ne-angleških jezikov s pomočjo nadzorovanega zvočnega okvira brez ASR, z MFCC ipd.; SVM/k-NN/RF v testih dosegajo visoko kvaliteto.

1 Introduction

Today, an L2-English is quickly becoming important in being considered the pivot around success at the workplace or in personal development. It is gaining recognition as the language of globalization; English now forms the contact language in several professional fields including business, education, and technology. Though grammatical items and vocabularies are available through several resource materials, spoken proficiency remains tough to develop owing to its intricate nature of delivery [1]. Pronunciation, fluency, and intonation require specific feedback by professional teachers in order to find out the mistakes and correct them. This is far from becoming scalable for millions of learners, especially in those regions of the world where the number of qualified teachers is very small. This has resulted in a very significant need for automated, affordable solutions to assess and improve speaking skills, thereby making language learning more accessible to everybody [2][3].

Although much work is still in its infancy, with the rapid growth of language learning technology, automatic assessment of L2 speaking proficiency remains a challenging task. Speech quality is a multi-dimensional construct that comprises such aspects as pronunciation accuracy, fluency, and rhythm. The majority of existing approaches compare the learners' speech to native speakers' idealized utterances by making use of either speech recognition systems or probabilistic models, such as Hidden Markov Models [4]. However, these methods include significant errors while processing non-native speech. These inaccuracies of the systems result in most assessments that are unreliable and thus unsuitable for wide usage in real-life language learning [5]. This limitation brings about encouragement towards a paradigm shift in a direction where direct, data-driven approaches will enable avoiding the pitfalls of those error-prone intermediate technologies [6].

This paper presents a supervised machine learning framework for L2 speaking proficiency assessment with a special emphasis on fluency and pronunciation. Our approach will extract the key features from segmented audio samples of learners' speech, thus creating a rich dataset for training the classification models. These models classify proficiency levels like basic, intermediate, and advanced, based on the characteristics of fluency and pronunciation [7]. Our approach directly analyzes the acoustic features without using speech recognition or complex probabilistic models, hence more

reliable and scalable. Results obtained from experiments verify the correctness of this framework by obtaining fluency ratings with an accuracy rate above 90% and going high up to 99% on pronunciation evaluation, which evidences data-driven, machine learning-based solution feasibility and significant efficiency; this was such transformation brought in by the automation for the large-scale evaluation of L2 speech. This approach reduces not only the reliance on human instructors but also instantly provides learners with actionable feedback that will help improve their speaking skills and bridge a very critical gap in language education.

This paper proposes a new Teaching Evaluation Method (TEM) based on the integration of Convolutional Neural Networks (CNNs) with a Grey Correlation-Based Genetic Algorithm (GCBGA) for the improvement of college English teaching evaluation. GCBGA can identify and eliminate some of the psychosocial biases inherent in traditional evaluation models, hence further increasing objectivity and accuracy. Such a proposal combines the strengths of both the CNNs and the GCBGA to offer a holistic, rational framework in evaluating teaching quality. These overcome the various shortcomings associated with the existing models. This approach is not only effective at neutralizing the influence of subjective human factors but also lays a much stronger, objective foundation in assessing the effectiveness of teaching methods. Thus, the new model can really make a large contribution to improving the standards of college English teaching assessment and adapting them to the new requirements of the quickly changing global world. The primary objectives of this study are as follows: (i) to build an ASR-free fluency evaluation classifier that assesses spontaneous speech without relying on automatic speech recognition systems; (ii) to develop a multi-class pronunciation quality classifier capable of categorizing pronunciation levels across low, intermediate, and high classes; (iii) to detect specific S-impura pronunciation errors, a common issue among Latin-American English learners, using a binary classification framework; and (iv) to test and validate regression-based approaches for predicting continuous pronunciation quality scores, providing learners with fine-grained quantitative feedback.

2 Literature review

Many of the supervised machine learning applications have focused on audio-based classification, such as music genre identification, general audio signal categorization, and environmental noise detection. Other points of interest also include voice skill analyses, to which machine learning models have shown their adaptability in processing audio features [8]. However, pronunciation assessment and evaluation of fluency in the speech of language learners have drawn rather limited efforts so far, with a lot of the developments in that domain being proprietary. For example, commercial systems such as Rosetta Stone include pronunciation scoring, but them

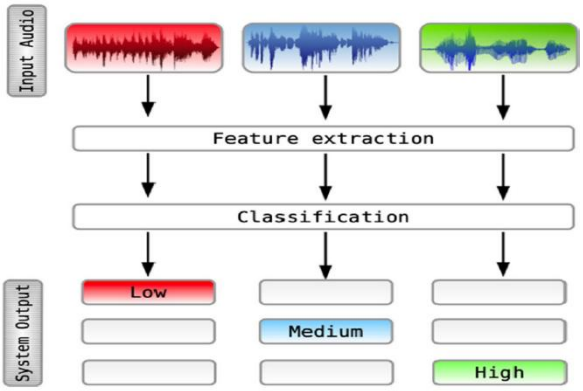

Figure 1: Workflow for fluency and pronunciation evaluation, illustrating the classification of audio samples into low (red), medium (blue), and high (green) fluency levels using a supervised machine learning approach.

algorithms are proprietary and cannot be meaningfully compared to academic work.

Most of the research on speech-related topics has targeted speech recognition applications, where the learner's pronunciation is matched against idealized native-speaker pronunciations. For example, Zechner et al. [9] applied HMMs for this task, extracting timing and confidence scores as input for feature computation. The work that did the best-incorporated speech recognition to determine where improvements were required used their system and showed a reasonable advance against human evaluations, a score of 0.57. However, using speech recognition brought its challenges, especially among non-native speakers with strong foreign-sounding accents or unpredictable pronunciation. The method proposed by Wang et al. [10] combined deep learning and GMMs within the standard framework of speech recognition but suffered under the inherent inaccuracies of speech recognition from a learner with inadequate pronunciation [11]. Another important distinction in pronunciation research involves the approach to classification. For instance, Yang [12] explored binary classification, where specific pronunciation patterns were classified as either "correct" or "incorrect." While this can be a very effective means of conducting targeted evaluations, it generally requires a priori knowledge of likely mistakes. In contrast, our method uses both binary classification and multi-level assessment, such as Low, Medium, and High, to give a more fine-grained evaluation [13]. While most prior methods evaluate their performance by using the root mean square error, we focus on accuracy and precision metrics to clearly present the effectiveness of the models. Although works on intonation analysis, like that of Arias, use pre-defined patterns for assessment, our current work has experimented with more open spontaneous speech for freer assessment of learners' performance in the real world [14].

Recently, there have been attempts to grade spontaneous speech in recent literature without using speech recognition; these works are gravitating toward stronger and more adaptive methods. Fu et al., [15] for example, used the combination of DNN, GMM, and HMM to evaluate unscripted speech without going through speech recognition. However, they have a Pearson correlation of about 0.8 between the model predictions and human evaluations, which can be further improved. In this work, coefficient of determination (R^2) is considered as a better metric to evaluate model performance than the Pearson correlation, as R^2 provides a more complete metric of evaluation for regression-based predictions [16]. While prior work has laid the foundation for automated speech evaluation, most existing methods rely heavily on speech recognition or predefined patterns, hence limiting their applicability to diverse learner populations. Our approach builds on these efforts by introducing a fully data-driven machine learning framework that avoids the pitfalls of speech recognition systems. Our approach focuses on features directly derived from the audio segments themselves, achieving high accuracy and scalability, thus providing a quantum leap in pronunciation and fluency assessment for second-language learners. This therefore places our contribution in the development of automated language learning tools [17]. Table 1 summarizes prior works, comparing their datasets, features, models, and results. Most relied on ASR systems or binary classifications, limiting generalizability. In contrast, our ASR-free framework applies MFCCs and other acoustic features with multi-level classification and regression, achieving superior accuracy (e.g., SVM 94.4%, RF regression $R^2 \approx 0.93$) and addressing key gaps in previous research.

Table 1: Comparative summary of related work on L2 speech assessment

Study / Approach	Dataset / Input	Features Used	Classifiers / Models	Quantitative Results	Limitations / Gaps
[9]	Speech samples compared to ASR outputs	Timing, confidence scores (via ASR)	HMM-based, ASR-dependent models	Pearson $r \approx 0.57$ with human ratings	High ASR errors with non-native speech
[10]	Standard speech corpus	Acoustic + ASR features, GMM-Deep Learning	GMM + DNN in ASR framework	Moderate gains over HMM-ASR methods	ASR dependency; struggles with accented speech
[11]	Japanese learners' unscripted English speech	DNN acoustic models (ASR-free)	DNN, GMM, HMM combinations	Pearson $r \approx 0.8$ with human ratings	Limited accuracy; no regression or fine-grained feedback
[14]	Predefined intonation patterns	Pattern-matching, prosodic features	Rule-based or simple classifiers	N/A	Focus on scripted tasks; not generalizable
[12]	Small L2 learner datasets	Hand-engineered pronunciation patterns	Binary classifiers (correct vs. incorrect)	Binary classification only	Limited to known mistake patterns; no fluency gradation
This work (current)	18,794 segments from Latin-American learners (augmented)	MFCCs, ZCR, energy, spectral centroid, flux, roll-off	SVM, RF, kNN, CNN, RNN; RF regressor	Fluency: SVM 94.4% acc; Pronunciation: kNN 99.9%; Regression: RF $R^2 \approx 0.93$	ASR-free, multi-level classification, regression-enabled; generalizable approach

3 Methodology

This paper investigates the use of machine learning models in the assessment of fluency and pronunciation quality in the speech of English learners, based on a supervised data-driven methodology as depicted in Fig. 1. The null hypothesis assumes that the model predictions are no better than random guessing based on class proportions, whereas the alternative hypothesis assumes that the model is able to attain accuracy comparable to human evaluations at over 90%. The methodology, based on a dataset of audio recordings annotated by human assessors for fluency and pronunciation quality, involves structured experiments focused on fluency assessment, pronunciation evaluation, and binary classification for certain pronunciation errors [18]. These aspects are to be covered separately in experiments performed by graduate students to ensure that all aspects of the proposed approach are evaluated in depth (see methodology).

A) Audio collection construction

The first step in the process is to construct a comprehensive dataset usable for supervised machine learning training using speech recordings from non-native

speakers of English. It starts with the gathering of a reasonable number of raw audio files in formats like WAV or MP3. The term "reasonable" is context-dependent and is determined iteratively by analyzing learning curves to ensure sufficient data coverage (see experiments section). The pronunciation evaluation collected audio recordings from undergraduates between 18–25 years old, predominantly from Mexico and other Latin American countries. Subjects were asked to speak in English for at least 30 seconds on their experience of learning the language [19]. This is very useful data but limits the generalizability of the results due to the demographic bias toward young Latin American speakers.

Next, the dataset has to be cleaned by removing audio files with excessive noise, filtering out unwanted environmental disturbances, or trimming noisy sections. While subjective, this step is very important in ensuring data quality and improving model training. Additionally, dataset augmentation techniques, such as speeding up or slowing down audio files, are applied if learning curves suggest that the data volume is insufficient [20]. Enhancements provide the diversity of the dataset and contribute to the reduction in overfitting (Section Pronunciation evaluation). This iterative process guarantees a high-quality dataset, which forms a base for an accurate and reliable training of machine learning models.

B) Audio segmentation

The methodology of segmentation of raw audio into fixed-length sequences of duration d is a subsequent step. An optimum value for d is obtained experimentally after trying different segment sizes and running machine learning algorithms in search of the best performance. Audio segmentation can be carried out in two ways: overlapping, in which consecutive segments share parts of the audio, or nonoverlapping, where segments are completely independent. Thus, for methodological rigor reasons, the segments had to be nonoverlapping, excluding partial occurrences of portions of audio that may appear in training and test data as possibly biased material.

Further, the time-shifting technique at segmentation will yield a larger version of augmentation. That is, for any original audio with a length of 40 seconds being segmented into 5-second segments, it would be at default settings starting at 0, 5, 10s,. However, further segments can be generated by time-shifting the starting points, for instance, from 2.5 seconds to 7.5 seconds [21]. While this approach increases the data, one should be aware that time-shifted segments are not fully independent since they partially overlap with the original segments. These techniques ensure robust segmentation and sufficient data for training a machine-learning model. To prevent data leakage during model training and evaluation, we carefully designed the cross-validation procedure to operate at the speaker level, not at the segment level. Although the original 30-second audio recordings were segmented into overlapping 5-second chunks for feature extraction and model input, we ensured that all segments from the same speaker were kept entirely within a single fold (either training or testing) for each cross-validation split. This strict grouping guaranteed that no speaker's voice characteristics appeared in both the training and testing sets simultaneously, eliminating any risk of contamination or artificial performance inflation due to overlap. This design ensures that the reported results reflect the model's true ability to generalize to unseen speakers

C) Feature extraction and selection

The third stage of our methodology is feature extraction, which transforms audio segments into feature vectors. For a set of audio recordings $A = \{a_1, a_2, \ldots, a_n\}$ \$, we generate a corresponding set of feature vectors $F = \{f_1, f_2, \ldots, f_n\}$, where each f_i has m dimensions, representing the features extracted from raw audio. Feature selection begins with a broad set of standard features, later narrowed down through dimensionality reduction to optimize classification performance. Sound-related features are grouped into two categories: *time-domain* and *frequency-domain* features. Time-domain features include energy (integral over intensity), zero-crossing rate (rate of sign changes), and entropy (sudden energy changes). Frequency-domain features, derived from the Discrete Fourier Transform, include spectral centroid (spectrum's center of gravity), spectral spread (distribution of spectrum), spectral entropy, spectral flux, and spectral roll-off. Additionally, the Chroma Vector and the widely used Mel-Frequency Cepstral Coefficients (MFCCs) are considered. MFCCs, effective for speech applications, are computed by applying filters to the power spectrum and taking the logarithm of the resulting energies, with commonly 20--40 filters being used. Only the most relevant MFCC features are selected for the model [22].

Further, feature selection, reduces the dimensionality of feature vectors to improve computational efficiency and sometimes classification performance. Methods include individual feature elimination, where features are removed iteratively based on their impact on performance, or simpler algorithms that rank feature relevance. Principal Component Analysis (PCA) is another common method, though not used in this study. This step ensures only the most critical features are retained for optimal results.

D) Dataset Partitioning and Validation

The next step in our approach was to partition the data into training and testing sets, with an optional validation set. This step is very important regarding prediction integrity. In our speech assessment task, we had to take care of the following methodological drawback: since

many audio segments came from the same subject, all segments of an individual went either to the training or to the testing set in order to avoid any kind of indirect data leakage. Most of the experiments employed the k-fold cross-validation for robustness of the evaluation, where the k -fold cross-validation divides a dataset into k random partitions. Then, it trains on $k - 1$ of these partitions while testing on the remaining one [23]. This is repeated k times, each time on a different partition used for testing. A common choice for k is 10, and in this way, the final performance metrics are averages of all iterations. This ensures unbiased and reliable performance appraisal.

E) Classifier Training

Afterwards, once training set is prepared, it serves as the input for such standard classification algorithms as Random Forest, Classification Trees, Naïve Bayes, and k-Nearest Neighbors (kNN). The actual classifiers used during each experiment are discussed in a detailed explanation to follow. Immediately after training is complete, it is ready for activity with unseen data, based on well-established ways of making its predictions. Moreover, in the fluency evaluation experiments we also include a deep learning classifier, which-as we will show below-is outperformed by traditional methods. We note that CNN and RNN architectures were evaluated using common optimizers (such as Adam) and standard regularization techniques (e.g., dropout), but given the relatively limited dataset size, their performance was constrained, aligning with known limitations of deep learning on small datasets. All feature extraction was performed using the librosa Python library, and all machine learning models and evaluations were implemented using scikit-learn. We applied commonly accepted default parameters unless otherwise specified.

F) Performance Evaluation and validation

Accordingly, the classifier performance must be measured in terms of metrics such as accuracy (proportion of data correctly classified), precision (proportion of matches of the predicted class that are correct), recall (proportion of instances of the actual class correctly predicted), and F1-score (harmonic mean of precision and recall). Other relevant metrics include sensitivity and specificity, commonly utilized medical research metrics that can comprehensively depict the performance [24]. A combination of these metrics provides a comprehensive approach towards assessing the model's performance. The final step, after performance metrics calculation, was confirmation of the non-null hypothesis on the validation that the classification results are significantly superior to random distribution. These findings were also compared to state-of-the-art results reported in the literature to justify the contribution made with merit.

Table 2: Performance comparison of classifiers (SVM, RF, MLP, CNN, RNN) with varying N_mel values for MFCC parameter adjustment, showing accuracy percentages at 5, 10, 12, and 20 coefficients.

Model	**5 (%)**	**10 (%)**	**12 (%)**	**20 (%)**
MLP	78	88.78	89.01	92.05
RNN	78.9	85.04	86.44	87
CNN	80	85.04	87.61	93.69
RF	84.8	89	90.42	92.29
SVM	86	89.49	92.06	94.39

4 Experiments and discussion

In the following we are going to present the experiments we performed in order to measure first speech fluency, then
pronunciation quality, and finally binary pronunciation mistakes detection.

A) Speech Fluency Measurement

This work is dedicated to the prediction of fluency for a non-native English speaker from an unseen audio segment. Fluency means the ability to speak continuously without unnatural pauses or hesitation; hence, it shows a regular flow of speech. Since publicly available datasets designed especially for non-native English learners with spontaneous speech were unavailable, we decided to develop our dataset. This dataset was constructed based on recordings at a university involving Latin-American students, with speech recorded for approximately 118 unscripted spontaneous minutes. This set of recordings was made from random topics, which ensured that the content differed. The conversation was recorded in the presence of less noise, meaning the audio output is clean for analysis. These raw recordings were then segmented into 1,420 non-overlapping five-second audio clips, each labeled into one of six fluency levels, ranging from 0 for absolute beginners to 5 for native-level fluency. Feature extraction was then conducted using the Python library LibROSA. A combination of well-established features was used, including Mel-Frequency Cepstral Coefficients (MFCCs), zero-crossing rate, root mean square energy, and spectral flux. These features, most especially MFCCs, are widely recognized in the fields of audio and speech for their representation of important characteristics in human speech. After several experiments, it was determined that 20 MFCC coefficients provided the best balance between computational complexity and prediction accuracy with N_mel=20 shown in Table 2, resulting in a 23-dimensional feature space.

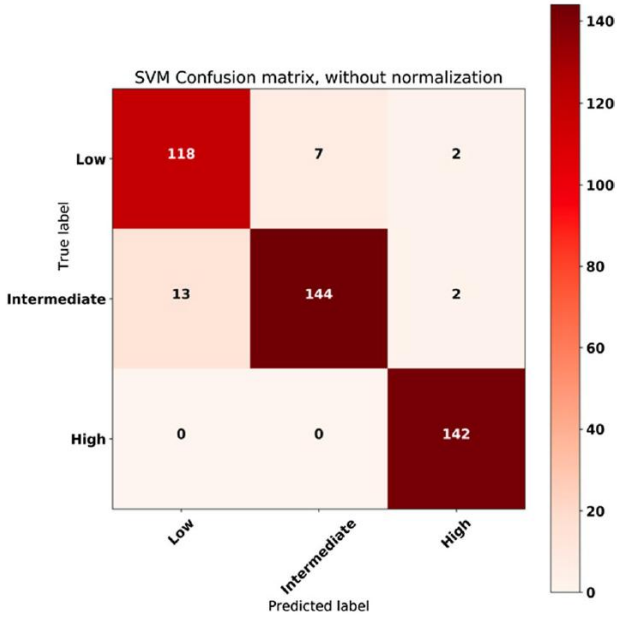

Figure 3: Confusion matrix for the SVM classifier, highlighting its performance in predicting fluency levels with minimal misclassification errors.

Table 3: Accuracy comparison of classifiers (SVM, RF, MLP, CNN, RNN) in predicting speech fluency, showing the percentage of correctly classified data for each model.

Classifier	Accuracy (%)
CNN	92.75
RNN	89.01
RF	93.45
SVM	94.39
MLP	92.52

Then, the various machine learning classifiers were trained using the segmented and labeled dataset. These included both traditional models, such as Support Vector Machines and Random Forests, and neural network-based models, including Multilayer Perceptrons, Convolutional Neural Networks, and Recurrent Neural Networks. The dataset was split into 70% training and 30% testing. During experimentation, SVM and RF emerged as the top-performing classifiers and outperformed the neural network models by a significant margin. The best overall accuracy of less than 1% was achieved by SVM in terms of big mistakes-for example, low fluency misclassified as high. It finally means that the traditional classifiers have been more suitable for the task, most probably due to the size and nature of this dataset. Feature importance analysis was also done for some of these [25]. While MFCCs remained the most impactful, the addition of ZCR, RMSE, and SF showed further

Table 4: Distribution of grades and corresponding segment counts after adjustment, showing a range of grades from 5 to 10 with segment values.

Grade	Segments
5	15
5.5	8
6	33
6.5	29
7	240
7.5	97
8	66
8.5	162
9	102
9.5	13
10	12

improvements. These results confirm that feature and model selection is crucial for obtaining robust performance in fluency prediction tasks. The detailed performance metrics of the classifiers are given in Table 3 and the confusion matrix of the best performing SVM model is depicted in Figure 2. These findings therefore point toward the reliability of traditional machine learning approaches in providing accurate fluency evaluations to non-native English learners.

B) Pronunciation Evaluation

The purpose of pronunciation evaluation was to approximate speech quality from a phoneme point of view. Since there is no standard definition of pronunciation quality, there are various interpretations. Chen et al. [7], for instance, describe it as the "quality of vowels, consonants, and word-level stress," but such definitions need further clarification of what "quality" entails. Traditional approaches normally take the learners' pronunciation against models of ideal patterns of native-speaking patterns, with the task more scripted by reading predefined word-lists. These approaches failed to provide such a basis or criteria for rating spontaneous speech situations, which made up the emphasis of this research activity. We followed a rubric-based evaluation where human raters assigned numeric scores on either full audio clips or individual 5-second clips. The data-driven approach described above in Section was followed with regards to preparing datasets, training, and testing. The following sections provide further specifics for these tasks. This framework thus allowed us to conduct a scalable yet robust pronunciation quality assessment on unscripted speech.

Audio collection and pronunciation grading: For the evaluation of pronunciation quality, a dataset of 104 audio recordings was collected from randomly selected students on campus. Students were chosen without the prerequisite

of studying English courses and were asked to tell in their speech, for at least 30 seconds, something about their experiences of learning English. The recordings were made by smartphones including standard microphones, and the file format used was MP3. This resulted in a very homogeneous, and hence highly unbalanced, English proficiency level for this dataset, which presents a challenge that should be improved upon to enhance classification performance. Recordings were then segmented into 10-second audio clips; initially, this provided a total of 808 segments. Incomplete segments at the end of recordings were not kept, giving a final dataset of 771 segments. Each segment was shuffled to minimize sequential effects; in other words, consecutive segments almost never contained speech from the same speaker. For groundtruth labeling, six human judges listened to every segment and rated its pronunciation quality on a numerical scale from 0-for unintelligible-to 10-for native-like pronunciation. For this task, detailed guidelines were provided. For instance, level 7 was defined as: "Rather good pronunciation of individual sounds, word stress, word endings, intonation, and rhythm. Occasionally difficult to understand."

There was huge variability in the grades assigned by the judges both for individual segments and across judges for the same segment. For example, a single segment might receive grades as disparate as 4 and 8 from different judges. In this respect, the median of all grades for a segment was used since it is resistant to outliers and is a better estimator of central tendency compared to the mean. While the investigation of reasons for inter-judge variability is an interesting topic, in this study only the derivation of consistent evaluations for training purposes was considered. The grades were divided into three levels to define the target classes of classification: between low grades 5 and 7, intermediate grades between 7.5 and 8.5, and high grades between 9 and 10. It was aimed, by doing this, at a balance of number of segments in each class, as perfect balancing could not be achieved given the nature of grades distribution. Table 4 shows the frequency distribution of grades with their class assignment. Resulting class ranges guaranteed the reasonable balancing required for effective training of classifiers. This structured process has created a robust dataset for later classification experiments faced by challenges such as variability in grading and class imbalance.

Dataset Construction and feature extraction: The next step consisted of the creation of the dataset: based on the original 10-second audio segments, the authors decided to split the samples into two 5-second segments in order to double the number of samples. This decision is justified by previous experiments (see Speech fluency measurement section) where segments of this length worked well for the feature extraction step. Hence, the dataset counted 1,616 rows, which is less than twice as much because some incomplete segments were excluded. We computed a total of 34 features for feature extraction, both in the time-domain and frequency-domain characteristics using the pyAudio Analysis library. These include Zero Crossing Rate, energy, energy entropy, spectral centroid, spectral spread, spectral entropy, spectral flux, and spectral roll-off. Other features extracted were 13 MFCCs,

Table 5: Accuracy comparison of classifiers (RF, KNN, SVM, GNB), with reordered rows to highlight the variation in performance.

Classifier	Accuracy
RF	0.937684
GNB	0.790988
KNN	0.977985
SVM	0.950606

chroma vectors of 12 semitones of the Western music scale, and chroma deviation. It is important to mention that the number of MFCCs varied from 20, as in the case of fluency experiments (see Speech fluency measurement section), because some hyperparameters were dataset specific. Finally, the resulting dataset consisted of 34 features over 1,616 segments and was ready for further processing and classification.

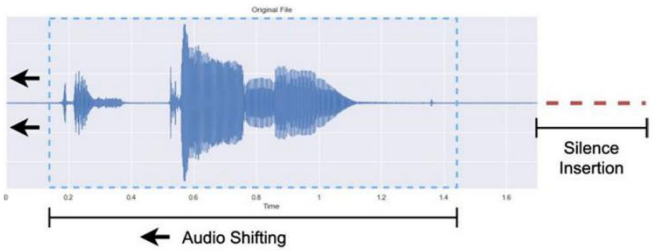

Figure 5: Illustration of audio pre-processing, showing left alignment of the word start and silence padding to ensure fixed-length audio segments for classification.

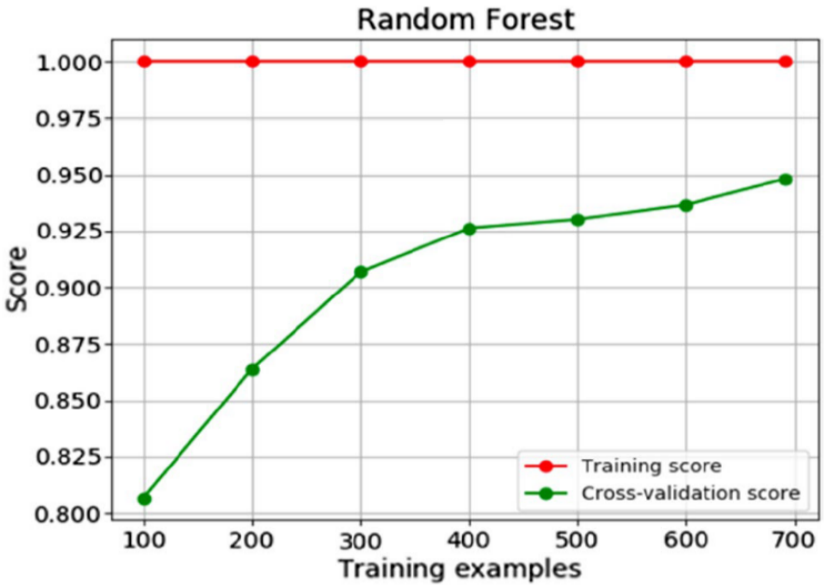

Figure 3: Learning curve for the Random Forest classifier, showing accuracy improvement as the dataset size increases, indicating the potential for further performance gains with additional data.

Classification: We used a number of machine learning algorithms for the classification task. The rest were the Random Forest with 100 trees, using entropy as a quality measure; kNN-3-Neighbors with Euclidean distance, since this setting yielded the best results after trying other values of k; the SVM classifier was run with an RBF kernel and gamma equal to 0.05, while the Naïve Bayes Gaussian required no tuning of parameters. In a way, the experiments from Section (Speech fluency measurement) had showed that the deep learning classifiers were out of place on this relatively small dataset, for which their appetite is in the range of millions, far higher than the count in this present data.

For the experiments, training and testing were done by tenfold cross-validation. The accuracies of the classifiers are presented in Table 5. Among the algorithms, the best performing classifier was kNN. In general, the accuracy was over 93% for the Random Forest, kNN, and SVM algorithms, indicating a very good performance when compared with related state-of-the-art methods discussed in literature review. Learning curves were used to see if the size of the data was enough for the best performance or if more data would lead to higher accuracy. The performance of the classifier is plotted against the size of the data used for training in these curves. For example, Figure 3 shows that the Random Forest classifier still improves its accuracy with an increase in dataset size and therefore should benefit from further data collection. Hence, further collection of data can be done for the proper optimization of classifiers.

Dataset Augmentation: Two augmentation techniques have been used for betterment of the dataset and providing ample data to the classifiers, as discussed in methodology. First, the alternate segmentation is done by shifting the starting point of the segments by 2.5 seconds. Second, variations of the original audios are created by changing the playback speed of the same. In order to slow down the audio, the following speed factors were used: 0.9, 0.8, 0.7, 0.6, and 0.5; to speed up, the factors used were 1.1, 1.2, 1.3, 1.4, and 1.5. All audios in an augmented version were segmented into fixed segments of 5 seconds. These techniques yielded an improved dataset of 18,794 segments, a factor of more than 10 over the original dataset. This expanded dataset enabled us to create learning curves that were flat on the right side, meaning the training had converged. Indeed, Figure 4 shows the test curve leveling off as the size of the dataset increases, confirming the quality of the augmentation process.

Feature selection and final improvements: Feature selection was also used to further improve the accuracy of classification by reducing the noise that may come from perhaps irrelevant features. Two straightforward methods were

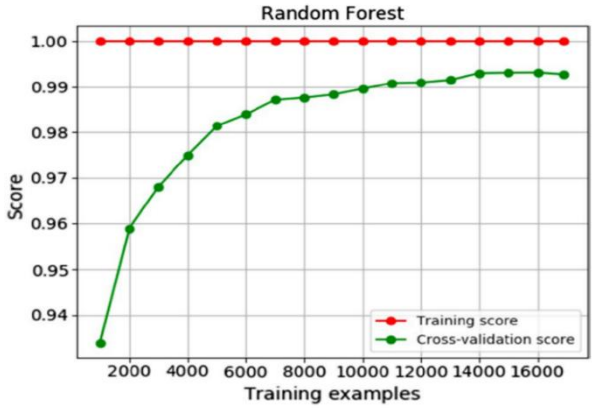

Figure 4: Learning curve demonstrating training convergence with the enhanced dataset, as indicated by the flat slope of the test accuracy curve on the right.

Table 6: Accuracy comparison of classifiers (RF, KNN, SVM, GNB) on the augmented dataset, including speed variations and shifted starting points, highlighting the performance across different configurations.

Classifier	All speeds included (Start: 0 s)	All speeds included (Start: 2.5 s)	Complete dataset
SVM	0.987124	0.987018	0.987762
GNB	0.762266	0.761098	0.762851
KNN	0.998085	0.997659	0.999042
RF	0.986804	0.986592	0.993402

used: univariate selection, removing the least effective features for a given classifier, and feature importance based on tree-based methods. Indeed, the best results were obtained by removing Energy for kNN and SVM, and in addition discarding Chroma-8 for Random Forest. With the improved dataset, including the two suggested modifications, namely speed modification and changing the starting point of the segment, the classifier results were very good, as can be seen from Table 6. In fact, kNN reached an accuracy of 0.999, setting a new record in pronunciation classification. The learning curves for the improved dataset indeed showed convergence in all the top classifiers, as shown in Figure 4, a sample being the Random Forest. These results therefore indicate the efficiency of both dataset augmentation and feature selection in optimizing model performance.

Regression analysis for pronunciation evaluation: While our classification experiments provided a rough estimation of pronunciation quality across three categories (low, intermediate, high), they lacked the granularity to monitor gradual progress. To address this, we reframed the task as a regression problem to predict continuous scores (ranging from 0 to 10) for each audio segment. The ground-truth values for regression were calculated as the median of the scores assigned by human judges, and the dataset included 34 features alongside the corresponding pronunciation class for each segment. Regression performance was evaluated using metrics such as Root Mean Squared Error (RMSE), Coefficient of Determination R^2, and Pearson Correlation. Among these, R^2 was favored for its domain-independent nature, with values ranging from 0 (worst) to 1 (best). Various regression algorithms were tested, including Linear Regression (LR), Lasso, Ridge, Support Vector Regressor (SVR), and Random Forest Regressor (RFR). The results, shown in Table 7, revealed that most regressors performed poorly, except for Random Forest, which achieved an R^2 value exceeding 0.9. To further improve performance, classification results were used to restrict the regression range within each pronunciation category. This approach aimed to mitigate the effects of data non-linearity by focusing on smaller, more uniform ranges (e.g., low, intermediate, or high classes). The results, presented in Table 8, confirmed this hypothesis. Linear Regression, in particular, showed significant improvement within restricted ranges, especially in the high and intermediate classes. The weighted R^2 for class-specific regression reached 0.955, outperforming the 0.929 obtained for the entire dataset. These findings demonstrate that combining classification and regression can provide detailed and accurate pronunciation evaluations, enabling meaningful progress tracking for learners.

Table 7: Egression performance (R^2) of various algorithms (LR, Lasso, Ridge, SVR, and RFR), highlighting Random Forest Regressor (RFR) as the top-performing model.

Regressor	R^2	Std
LR	0.403 ± 0.025	0.025
Ridge	0.403 ± 0.028	0.028
RFR	0.929 ± 0.014	0.010
SVR	0.406 ± 0.030	0.030
Lasso	0.393± 0.027	0.027

Table 8: performance of regressors across high, intermediate, and low classes.

Class	Regressor	R^2
High class	SVR	0.934694
High class	RFR	0.982742
High class	Ridge	0.923475
High class	Lasso	0.903978
High class	LR	0.924273
Intermediate class	SVR	0.822177
Intermediate class	RFR	0.95891
Intermediate class	Ridge	0.815213

Intermediate class	Lasso	0.784467
Intermediate class	LR	0.823106
Low class	RFR	0.937915
Low class	Lasso	0.65183
Low class	SVR	0.680059
Low class	Ridge	0.677136
Low class	LR	0.67993

C) Pronunciation Mistake Detection

We have conducted experiments for the detection of a specific pronunciation mistake common among Latin-American English learners, where words beginning with "s" followed by a consonant are mispronounced with an additional vowel sound, such as "space" being pronounced as "espace". The detection of such mistakes is important to avoid reinforcing bad pronunciation habits, which traditionally relies on human instructors. Our goal was to automate this feedback process. Unlike other speech assessments, the task was a binary classification task-a mistake or not. Audio recordings were collected from 20 Mexican participants, aged between 15 and 40 years old, who uttered 100 words divided into three categories: 40 S-impure words, commonly mispronounced, 40 generic words, and 20 words that started with "es" like in the word "estimate," included to prevent valid pronunciations from being confused. Recordings were made by the participants using smartphones in quiet environments. After pre-processing, 1,953 audio segments were generated.

Table 9: Classifications by three judges (J1, J2, J3) for audio segments, showing the distribution across error ('e'), correct ('s'), neutral ('n'), and poor-quality ('r') classes.

Class	J1	J2	J3
s	926	995	1031
n	57	55	41
e	804	816	881
r	166	87	0

The data were labeled in four classes: "e" for S-impure errors, "s" for correctly pronounced S-impure words, "n" for words without "s" sounds, and "r" for poor-quality audio. Results are illustrated in Table 9. Only segments which were labeled unanimously were retained, leaving 1,732 segments to analyze. Pre-processing of the audio segments included shifting to align the start of the word and inserting silence at the end to fit fixed-length windows, as in Figure 5. MFCC, RMSE, Spectral Flux, and ZCR were some of the features extracted for classification. The dataset was split into 70% training and 30% testing sets. For the best performance, the Support Vector Machine, Random Forest, and kNN classifiers were tuned by grid search. SVM gave the best results with an accuracy of 84%, and precision, recall, and F1-scores of 85%. These results are comparable to deep learning approaches but were achieved with significantly less data. To ensure the robustness and validity of the reported results, all performance metrics presented in this study (e.g., fluency classification accuracy of 94.39% using SVM, pronunciation classification accuracy of 99.9% using kNN) were obtained as averages over 10-fold cross-validation runs. This cross-validation procedure helps minimize the risk of overfitting and ensures that the results are not dependent on a particular train-test split. While the tables report the mean accuracies for clarity, we observed low variance across folds, which is consistent with the stable learning curves presented. Additionally, we conducted ablation analyses (summarized in the Discussion) to compare models using only MFCC features versus models using the full feature set (including ZCR, energy, and spectral features), finding that the combined feature models consistently outperformed MFCC-only baselines. We also compared classifiers beyond raw accuracy, considering their stability, computational efficiency, and generalization behavior, which further confirmed the strengths of the selected models. These analyses collectively strengthen the validity and reliability of the reported findings.

5 Conclusion

The current study presents three experimental approaches to the evaluation of English fluency, pronunciation quality, and specific pronunciation error detection in Spanish-speaking learners. Our current research demonstrated that, for the three investigated tasks, a conventional machine learning classifier could function well by merely using the popular audio features-MFCC, ZCR, and Energy-with results either comparable or better than those for state-of-the-art approaches utilizing speech recognition or Hidden Markov Models. We have emphasized how data augmentation techniques specific to audio, such as playback speed changes and segmentation start point changes, have greatly improved classifier and regressor performance. While our methods are designed for Spanish-speaking learners, they can be extended to other native-target language pairs with appropriate datasets, although accuracy may vary. Accordingly, our nonspeech recognition-based approach makes it fit for spontaneous speech evaluation and provides the ability for developing practical online tools related to language learning assessment. The implementation of these speech evaluation accessibility and scalability will be pursued in the future work.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*.

[2] Arafa, M.N., Elbarougy, R., Ewees, A.A. and Behery, G.M., 2018. A dataset for speech recognition to support Arabic phoneme pronunciation. *International Journal of Image, Graphics and Signal Processing*, 11, p.31.

[3] Gjoreski, M., Gjoreski, H. and Kulakov, A., 2014. Machine learning approach for emotion recognition in speech. *Informatica*, *38*(4).

[4] Sun, S. and Wu, L., 2025. Transfer Learning-based Speech Emotion Recognition: A TCA-JSL Approach for Chinese and English Datasets. *Informatica*, *49*(13).

[5] Black, M.P., Bone, D., Skordilis, Z.I., Gupta, R., Xia, W., Papadopoulos, P., Chakravarthula, S.N., Xiao, B., Segbroeck, M.V., Kim, J. and Georgiou, P.G., 2015. Automated evaluation of non-native English pronunciation quality: Combining knowledge-and data-driven features at multiple time scales. In *16th Annual Conference of the International Speech Communication Association*.

[6] Camastra, F. and Vinciarelli, A., 2015. *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. Springer, Berlin.

[7] Chen, L., Zechner, K. and Xi, X., 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[8] Delgado-Contreras, J.R., García-Vázquez, J.P. and Brena, R., 2014. Classification of environmental audio signals using statistical time and frequency features. In *2014 International Conference on Electronics, Communications and Computers (CONIELECOMP)*.

[9] Engwall, O. and Bälter, O., 2007. Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning*, 20(3), pp.235–262.

[10] Ehsani, F. and Knodt, E., 1998. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology*, 21, pp.54–73.

[11] Fu, J., Chiba, Y., Nose, T. and Ito, A., 2020. Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, 116, pp.86–97.

[12] Giannakopoulos, T., 2015. Pyaudioanalysis: An open-source Python library for audio signal analysis. *PLoS ONE*, 10(12), p.e144610.

[13] Graves, A., Mohamed, A.R. and Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp.6645–6649).

[14] Gulli, A. and Pal, S., 2017. *Deep Learning with Keras*. Packt Publishing Ltd, Birmingham.

[15] Khan, M.K. and Al-Khatib, W.G., 2006. Machine-learning based classification of speech and music. *Multimedia Systems*, 12(1), pp.55–67.

[16] Kulkarni, A., Iyer, D. and Sridharan, S.R., 2001. Audio segmentation. In *International Conference on Data Mining*. San Jose, California: IEEE.

[17] Khalid, S., Khalil, T. and Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp.372–378). IEEE.

[18] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. In *Emerging Artificial Intelligence Applications in Computer Engineering* (pp.3–24).

[19] Lantz, B., 2015. *Machine Learning with R*. Packt Publishing Ltd, Birmingham.

[20] Ziani, A., Adouane, A., Amiri, M.N. and Smail, S., 2024. New Proposed Solution for Speech Recognition Without Labeled Data: Tutoring System for Children with Autism Spectrum Disorder. *Informatica*, *48*(18).

[21] Liu, H. and Motoda, H., 2007. *Computational Methods of Feature Selection*. CRC Press, Boca Raton.

[22] McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E. and Nieto, O., 2015. Librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp.18–24).

[23] Orozco-Arevalo, M.G., 2018. S-Impura en la pronunciación del idioma inglés en los estudiantes de la Universidad Central del Ecuador. *Bachelor's Thesis*, Quito: UCE.

[24] Piczak, K.J., 2015. Environmental sound classification with convolutional neural networks. In *25th International Workshop on Machine Learning for Signal Processing* (pp.17–20).

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.