MACHINE LEARNING APPROACHES TO IMPROVING PRONUNCIATION ERROR DETECTION ON AN IMBALANCED CORPUS

Xuesong Yang[†], *Anastassia Loukina*[‡], *Keelan Evanini*[‡]

[†]Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, Urbana, IL 61801
[‡]Educational Testing Service, Princeton, NJ 08541
xyang45@illinois.edu, {aloukina,kevanini}@ets.org

ABSTRACT

In this paper, we investigate the task of phone-level pronunciation error detection as a binary classification problem, the performance of which is heavily affected by the imbalanced distribution of the classes in a manually annotated data set of non-native English. In order to address problems caused by this extreme class imbalance, methods for cost-sensitive learning (weighting inversely proportional to class frequencies) and over-sampling of synthetic instances (SMOTE) are investigated in order to improve classification performance. Experiments using classifiers consisting of features based on acoustic phonetics and word identity demonstrate that these machine learning approaches lead to performance improvements over the baseline system based on the extremely imbalanced data. In addition, several different types of classifiers were compared. Finally, the paper analyzes the robustness of classifier performance across different phones.

Index Terms: Imbalanced Learning, Sampling Methods, Pronunciation Error Detection, Spoken Language Assessment

1. INTRODUCTION

Computer assisted pronunciation training (CAPT) systems have attracted considerable attention from the speech and applied linguistics research communities in recent years. The goal of CAPT systems is to enhance non-native learners' spoken language skills in a foreign language by providing both an accurate assessment of the learner's pronunciation proficiency as well as diagnostic feedback to aid in learning.

Most previous work has focused on error detection in the context of pronunciation training systems; the goal of such systems would be to identify persistent errors in a non-native speaker's speech and suggest directions for further training. In this paper, on the other hand, we consider the task of pronunciation error detection in the context of large-scale assessment of English proficiency. While both assessment and training ultimately pursue similar aims, there are several restrictions posed by language assessment that need to be taken into account while designing the error detection system. For example, fairness considerations require that the system should not apply different criteria to test-takers with different native languages (L1). This means that the assessment system may not use the information about the test-taker's L1 to determine prior probabilities of error patterns as is frequently done in pronunciation training systems (e.g., [1]).

For our system, we cast the task of pronunciation error detection as a binary classification problem based on a set of features consisting of acoustic information and word identity. In nearly all cases, the number of phones labeled as pronunciation errors in the corpus is very small in comparison to the number of phones labeled as correct; this heavily skewed class distribution leads to challenges in modeling and evaluation. We investigate two common approaches (cost-sensitive learning and sampling) to mitigate problems caused by the extremely imbalanced distributions. We also analyze the robustness of classifiers across different phones and discuss which classifiers would be most effective in the context of a practical CAPT or language assessment system.

2. RELATED WORK

2.1. Pronunciation error detection

Recent work on pronunciation error detection has approached the task as a supervised learning problem. There are two specific types of implementation for this general approach.

One common approach is to utilize prior knowledge of mispronunciation patterns extracted from a large corpus of second language (L2) speech. Based on rules and statistical generalizations extracted from these mispronunciation patterns, the system's pronunciation dictionary can be extended to include mispronunciations, and prior probabilities can be added for the variants. The pronunciation error detection task then consists of identifying which realization of a given word occurred in a speaker's response. This can be done by either using algorithms built into the automatic speech recognition (ASR) engine to select the most probable variant [1] or by training a classifier based on various acoustic parame-

ters which are likely to differ between predicted realizations [2, 3]. This approach produces the best results when the predicted error patterns are customized to a specific combination of L1 and L2. However, as discussed above, such a customization may not be possible or desirable in the case of a global, large-scale language assessment with many diverse L1 populations.

Another line of research is to identify pronunciation errors based on the similarity between a speaker's realization of a given phone and its target pronunciation based on a native-speaker acoustic model. In many cases, the similarity metrics are based on ASR confidence scores; for instance, the most widely used metric, the Goodness of Pronunciation (GOP) score [4], calculates the duration-normalized log of the posterior probability that a speaker uttered a specific phone given the acoustic observations.

While earlier work on GOP [1, 4] aimed to establish appropriate thresholds for different phones and speakers, subsequent studies have re-cast it as a classification task in which GOP-like measures are combined with additional acoustic and supersegmental features using different machine learning algorithms. Some of the machine learning methods used in previous studies include logistic regression [5], linear discriminant analysis [2], support vector machines [6], and decision trees [7]. These studies mostly focused on comparing the performance of the same machine learning algorithm on different feature sets, and seldom provided insights into how to choose the most appropriate classifier for the specific task. Yet other work in the field of machine learning has demonstrated that the choice of which classifier to use may have a substantial effects on the system's performance due to task-specific strengths and weaknesses of various classifiers [8]. In this paper, we therefore compare different machine learning algorithms using the same set of features to evaluate whether and to what extent the choice of classifier may affect the performance of the pronunciation error detection models.

2.2. The problem of imbalanced learning

Classification problems involving real-world data are often imbalanced, and have a highly skewed distribution of classes. Despite this, most standard learning algorithms assume a balanced class distribution or equal misclassification costs. Once such algorithms are applied to (extremely) imbalanced data sets, the false acceptance rate tends to increase, because the model does not adequately estimate the distribution of the classes [9]. Two common approaches that have been investigated to address the problems caused by extremely imbalanced data sets are cost-sensitive learning and sampling methods [9]. Cost-sensitive learning methods assign a relatively high cost to misclassifications of minority class instances and minimize the overall cost, while sampling methods attempt to balance the class distributions by adjusting the relative proportion of instances in the distribution. These sampling-based methods, however, have some potential drawbacks; specifically, under-sampling (removing instances from the majority class) may cause the classifier to ignore important information pertaining to the majority class, whereas an over-sampling approach (duplicating instances from the minority class) may cause "tied" issues and lead to overfitting [10]. The synthetic minority over-sampling technique (SMOTE) [11] has been proposed to overcome these issues with sampling methods by generating artificial data based on similarities in the feature space across instances in the minority class.

Pronunciation error detection, cast as a binary classification task, also faces the standard problems that are caused by an imbalanced distribution of classes in the corpus. The amount of phones which are labelled as errors by expert annotators is extremely small in relation to the overall number of phones. For example, [5] notes that they were not able to train classifiers well for a number of phones because of the low percentage of annotated errors for those phones (cf. also [12]). To address this problem, we investigate the two previously mentioned approaches to deal with imbalanced data sets, and compare the performance of assigning weights inversely proportional to the class frequencies (Auto-Weighting) and the SMOTE over-sampling methods across a range of different classifiers. In the end, we also empirically evaluate the quality of different classification decisions and suggest the most effective classifier for the task of identifying mispronounced phones.

3. DESCRIPTION OF THE DATA

3.1. Corpus

The corpus of spoken responses used in this study was collected during the pilot administration of an international assessment of English proficiency targeted at middle-school students aged from 11 to 15. It consisted of 175 responses obtained from 175 native speakers of Korean, Arabic, Spanish and Vietnamese. All speakers were learners of English as a foreign language and resided in non-English speaking countries. Each speaker was asked to read one of the four texts out loud (45 responses for each text). The responses were manually transcribed and were automatically aligned with human transcriptions using the HTK-based Penn forced aligner [13].

This corpus was annotated for pronunciation errors by two linguists who annotated pronunciation errors following the approach from [14] in which raters were asked to identify "The most serious errors to be corrected in the subject's speech." The annotators used their own judgment about what errors should fall under this category; they were provided with the phonetic dictionary transcriptions of each word and were asked to modify the transcriptions for errors that they considered serious enough to break communication. For each text, six files were selected for double annotation to test inter-annotator agreement. The remaining files were split between

the two annotators using stratified sampling so that each annotator was assigned a similar number of responses for each L1. The files selected for double annotation were interspersed with the other responses, and the annotators were not aware which responses were selected for double annotation.

3.2. Human inter-annotator agreement

On average, the annotators corrected about 7% of all phones. This number varied widely among the phones: for example, 29% of all occurrences of /DH/ were marked as mispronounced, while for /M/ this number was just 0.01%.

For the doubly annotated responses, we aligned the transcriptions using edit distance and computed the absolute agreement (% of matching values) and Cohen's kappa (κ) on the phone level for each response. The inter-annotator agreement on the localization of errors varied between items with an average $\kappa = 0.52$ and an average absolute agreement 92%. In addition, the two annotators agreed strongly on the relative number of mispronounced phones in each response with Pearson's $r = 0.90$ ($p = 3 \times 10^{-6}$, $N = 24$) for the number of phones corrected by each annotator per response.

These results compare favorably with inter-annotator agreement results reported in previous studies. For example, [7] reported 80.2% agreement for the localization of phone-level pronunciation errors in a corpus of Spanish. For English, [15] reported 67% agreement on the localization of phone-level errors. [16] reported Intraclass Correlation Coefficient (ICC) values between 0.29 and -0.56. The annotation procedure used in this study consistently produced agreement above these reported values.

To evaluate the validity of our annotations, we computed correlations between the number of words corrected by the annotator and the holistic proficiency score assigned by the first human rater (this holistic score was based on an evaluation of several aspects of English speaking proficiency, including delivery, vocabulary, grammar, and content, and was not limited solely to an evaluation of pronunciation). For responses that were annotated by both annotators we used the mean value of the number of corrections from the two annotations. The overall correlation between the number of corrections and the holistic proficiency score was $r = -0.57$ ($p = 3.02 \times 10^{-22}$, $N = 175$).

4. METHODOLOGY

Automated scoring of non-native speech relies on automatic speech recognition to convert the spoken response to a text transcription. For this study, we used a state-of-the art ASR system to recognize the speech from the target corpus with acoustic models trained on non-native speech and adapted to both children's speech and in-domain data [17].

For this study, we only used the words where the ASR hypothesis was in agreement with the human transcription (14,302 words out of 20,772 words). This was done to ensure that our system learns to identify actual mispronunciations rather than ASR errors. This procedure inevitably led to some mispronunciations being excluded from the analysis; these instances will be analyzed in a subsequent study. The final corpus consisted of 50,261 phones (error: 3,665, correct: 46,596).

4.1. Acoustic features and word identity

For each phone, the six acoustic features listed in Table 1 were extracted. Acoustic likelihood scores correspond to raw likelihoods. Confidence scores are raw posterior probabilities computed based on the phone lattice. For each measure we also used a duration-normalized version computed by dividing the raw or posterior probability for each phone by the number of frames in the phone.

Each speaker in this study read one of four texts. Preliminary analysis of human annotations showed that some words were more likely to contain errors than others, and similar patterns have also been observed in previous work on pronunciation error detection [16]. For instance, 75% of speakers mispronounced the word 'barley'. To ensure that our models identify actual mispronunciations rather than simply learn difficult words, we also trained models using a word identity feature and used these models as a baseline. Finally, we trained models which combined both the word identity feature and the acoustic features. All models are summarized in Table 1.

Table 1: *Models used in this study*

Model	Feature	Description
`ac`	*am*	Acoustic likelihood
	am_dur	*am* divided by N frames
	cs	Posterior probability of each phone
	cs_dur	*cs* divided by N frames
	logcs	Log of *cs*
	logcs_dur	*cs* divided by N frames
`wi`	word identity	
`ac+wi`	`ac` and `wi`	

4.2. Classifiers and imbalanced learning

We trained and evaluated separate models for each of the 39 phones, since previous work has found that the distribution of acoustic features differs across phones (see, for example, [4]). Six classifiers[1] were selected to distinguish pronunciation errors on the imbalanced corpus: decision trees, random forest, gradient boosting, support vector machines with a linear kernel (LinearSVC) and a radial basis function kernel (SVC), and binomial logistic regression.

To address the issues caused by the imbalanced distribution of classes in the corpus, we experimented with the two approaches described in Section 2.2. First, we increased the

[1]SKLL package: https://github.com/EducationalTestingService/skll

cost of misclassifying instances from the minority class by assigning weights inversely proportional to class frequencies (Auto-Weighting). Due to practical limitations, this procedure was only applied for three classifiers: SVC, Linear SVC and Logistic Regression. We also applied the synthetic minority oversampling method discussed in Section 2.2 (SMOTE). This was done for all 6 classifiers.

The models were evaluated using 10-fold stratified cross-validation. For the balanced data sets, the oversampling was applied only to the folds in the training set and not to the fold used for evaluation. We use repeated measures ANOVA to compare the performance of the classifiers and models within each phone.

5. RESULTS

5.1. Imbalanced data

Figure 1(a) shows the distribution of *F-score* over all classifiers for each phone for `ac` model. Red dots indicate the average performance for the baseline `wi` model over all classifiers. For the two phones *AA* and *Y*, the baseline `wi` achieved an *F-score* greater than 0.55. In most other cases the baseline model performed at chance level, whereas the `ac` model generally performed above chance level. However, for four phones (*P*, *AH*, *AA* and *Y*), the `ac` model did not outperform the baseline `wi`. The performance for most phones remained relatively low; the *F-score* was below 0.1 for 32 out of the 39 phones with an average *F-score* of 0.09 across all phones. (see Table 2).

Table 2: *Mean F-score value for different models (columns) and methods of balancing the data (rows). SMOTE-3 includes the three classifiers where we also applied auto-weighting. SMOTE-6 includes all six classifiers.*

	wi	*ac*	*ac+wi*
Imbalanced	0.05	0.09	0.16
SMOTE-6	0.22	0.20	0.23
SMOTE-3	0.22	0.22	0.25
Auto-Weight	0.21	0.24	**0.27**

Combining together both acoustic features and word-identity `ac+wi` leads to a further increase in *F-scores* as shown in Figure 1(b). The performance across all phones, however, remains low, with an average *F-score* of 0.16.

We also observed that the average *F-score* was correlated with the percentage of errors for each phone: r=0.509 (p <0.001). In other words, the models performed better for phones with a larger percentage of errors, thus confirming our initial intuition that the imbalanced nature of the data set may affect the performance of the model.

5.2. Balanced data

We next conducted a study on balancing the training corpus by the Auto-Weighting and SMOTE approaches to improve the classification performance based on the combined `ac+wi` model compared to `wi` baseline.

5.2.1. Auto-Weighting and SMOTE

Figure 2 shows the ranked *F-score* over all phones and classifiers for different combinations of features and imbalanced learning approaches. We see large increases in performance for the baseline `wi` model when the two approaches are applied, and moderate increases in performance for the `ac+wi` model in comparison to the original imbalanced data set (p = 6×10^{-7} for model, p = 2×10^{-16} for sampling method, p = 2×10^{-16} for interaction). Table 2 shows that the mean *F-score* value increased from 0.16 on the imbalanced data set to 0.27 using the Auto-Weighting approach.

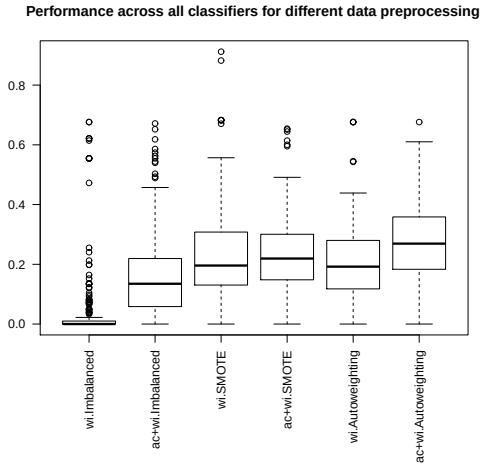
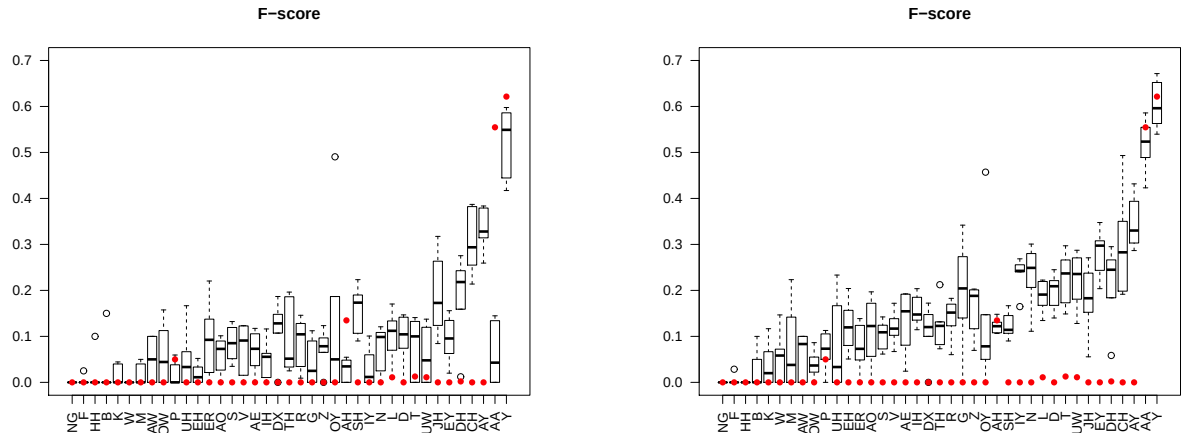


Fig. 2: Performance across all classifiers and all phones

5.2.2. Comparison between classifiers

We further explored the difference among the six classifiers by comparing the average *F-score* for the combined model `ac+wi` for each classifier over all phones. Figure 3 illustrates that decision trees and a support vector machine with a non-linear kernel (SVC) outperformed the two linear classifiers (logistic regression and linear SVC) on imbalanced data, but linear classifiers outperformed the other classifiers after applying the two balanced learning approaches (the classifier effect on *F-score* within each phone: p = 0.00162 after controlling for model and oversampling method).

Figure 4 illustrates the performance of the support vector classifier with a linear kernel using two different imbalanced learning approaches across different phones for both the `wi` baseline and its improved model `ac+wi`. It shows that both approaches for the combined model `ac` + `wi` achieved similar *F-scores* and outperformed the baseline `wi` model for most phones. The performance of classifiers varied substantially across phones.



(a) Imbalanced: `ac` v.s. `wi`

(b) Imbalanced `ac+wi` v.s. `wi`

Fig. 1: Classifier performance for different models on the original data. The boxes show *F-score* across different classifiers for each phone for a given model (`ac` or `ac+wi`). Red dots indicate the mean *F-score* across all classifiers for a baseline model based on word-identity only.

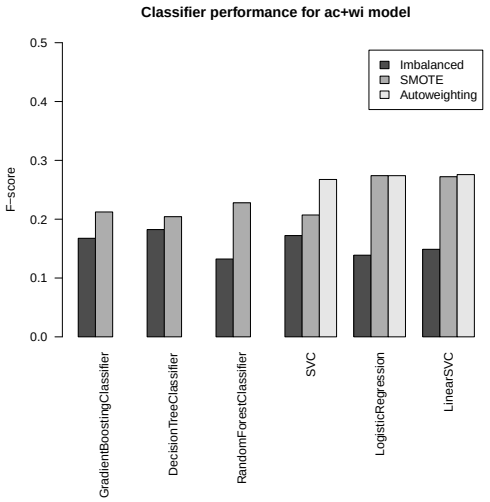


Fig. 3: Classifier performance for the `ac+wi` model

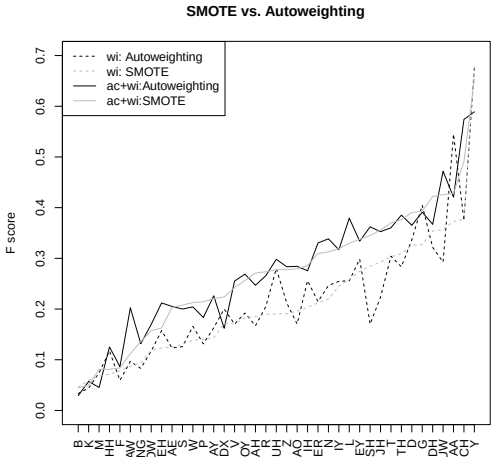


Fig. 4: Comparison of Auto-Weighting and SMOTE imbalanced learning approaches for support vector machine with linear kernel.

6. DISCUSSION

In this study, we compared the performance of different machine learning algorithms and approaches to handling imbalanced data sets in the context of the task of pronunciation error detection in a large-scale language assessment. We found that the best performance could be achieved by combining information about word identity and the acoustic properties of the word, although the performance of the models varied across phones. We note that a simple model based solely on the word identity achieved relatively high performance for some of the phones, especially on the balanced data set (for *AA*, *F-score* = 0.55). Despite this high performance, such a model has little use in assessment or training: it only distinguishes difficult and easy words without regard to the correctness of a particular pronunciation. Nonetheless, this potential effect of word identity has been often ignored in previous studies. We recommend that a model based on word identity should be used as one of the baselines in all future studies on pronunciation error detection to ensure that the model performance is not limited to the identification of difficult words.

Both cost-sensitive (Auto-Weighting) and synthetic minority (SMOTE) approaches substantially improved the performance of the model in comparison to the baseline trained on the original imbalanced data. For some classifiers, the Auto-Weighting approach outperformed SMOTE. Since the Auto-Weighting approach only considers the cost associated with misclassifying samples, while the SMOTE approach synthesizes artificial data, the lower performance of the SMOTE method may be due to the fact that the synthetic data was generated from only the minority class and may have led to an increased overlap between classes.

Finally, we found that support vector classifiers and logistic regression obtained better classification performance than decision trees, random forests, and gradient boosting classi-

fiers. We also observed that classification performance differed by phone and, in future work, we plan to explore possible linguistic factors contributing to this difference.

7. CONCLUSION

This paper investigates the task of phone-level pronunciation error detection as a binary classification problem, of which the performance is highly affected by the imbalanced distribution of classes (error vs. correct) in the data set. We explored the use of a word identity feature as a baseline, and acoustic features combined with it improved the overall classification performance on both imbalanced and balanced data. Meanwhile, two imbalanced learning approaches (Auto-Weighting and SMOTE) were applied and both achieved better average *F-scores* by a large margin. In the end, empirical experiments were also performed for different machine learning classifiers, among which support vector machines with a linear kernel and logistic regression were the most effective classifiers on the general task of identifying pronunciation errors.

8. ACKNOWLEDGMENTS

We thank Nils Ever Murrugarra Llerena for many useful discussions about this project and also thank Daniel Blanchard and Nitin Madnani for their help with the SKLL toolkit, Melissa Lopez and Hillary Molloy for annotating the data.

References

[1] Dean Luo, Xuesong Yang, and Lan Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus.," in *Proceedings of Interspeech*, 2011, pp. 1593–1596.

[2] Helmer Strik, Khiet Truong, Febe de Wet, and Catia Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[3] Carlos Molina, Nstor Becerra Yoma, Jorge Wuth, and Hiram Vivanco, "ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion," *Speech Communication*, vol. 51, no. 6, pp. 485–498, 2009.

[4] Silke M. Witt and Steve J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[5] Joost van Doremalen, Catia Cucchiarini, Helmer Strik, and Joost Van Doremalen, "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1336–1347, 2013.

[6] Si Wei, Guoping Hu, Yu Hu, and Ren-Hua Wang, "A new method for mispronunciation detection using Support Vector Machine based on Pronunciation Space Models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[7] Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash, and Kristin Precoda, "Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[8] Rich Caruana and Alexandru Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 161–168, 2006.

[9] Haibo He and Edwardo A Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[10] David Mease, Abraham J Wyner, and Andreas Buja, "Boosted classification trees and class probability/quantile estimation," *The Journal of Machine Learning Research*, vol. 8, pp. 409–439, 2007.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and Philip W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] Nancy F. Chen, Vivaek Shivakumar, Mahesh Harikumar, Bin Ma, and Haizhou Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages.," in *Proceedings of Interspeech 2013, Lyon, France*, 2013.

[13] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3878, 2008.

[14] Ambra Neri, Catia Cucchiarini, and Helmer Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training," *IRAL*, vol. 44, no. 4, pp. 357–404, 2006.

[15] Patrizia Bonaventura, Peter Howarth, and Wolfgang Menzel, "Phonetic Annotation of a Non-Native Speech Corpus," in *InSTIL 2000 (Integrating Speech Technology in (Language) Learning), pp. 10-17, Dundee, UK, 29-30 August*, 2000, pp. 225–230.

[16] Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Nöth, and Satoshi Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[17] Keelan Evanini and Xinhao Wang, "Automated speech scoring for non-native middle school students with multiple task types," in *Proceedings of Interspeech 2013, Lyon, France*, 2013, pp. 2435–2439.