# Applying Feature Bagging for More Accurate and Robust Automated Speaking Assessment

Lei Chen

*Educational Testing Service*
*Princeton NJ USA*

lchen@ets.org

***Abstract*—The scoring model used in automated speaking assessment systems is critical for achieving accurate and robust scoring of speaking skills automatically. In the automated speaking assessment research field, using a single classifier model is still a dominant approach. However, ensemble learning, which relies on a committee of classifiers to predict jointly (to overcome each individual classifier's weakness) has been actively advocated by the machine learning researchers and widely used in many machine learning tasks. In this paper, we investigated applying a special ensemble learning method, feature-bagging, on the task of automatically scoring non-native spontaneous speech. Our experiments show that this method is superior to the method of using a single classifier in terms of scoring accuracy and the robustness to cope with possible feature variations.**

**Index Terms**: speech assessment, ensemble learning, feature bagging, speech recognition

## I. Introduction

In the last decade, a large number of studies have been conducted using automated speech recognition (ASR) technology to automatically score speech or evaluate pronunciation ( refer a comprehensive review [1]). In these automated speech assessment systems, different speech features were computed using various methods, e.g., signal processing, prosody analysis, and natural language processing (NLP). The extracted features were fed into a statistical model to automatically predict human speaking proficiency levels. Several types of classifiers from the machine learning research were used. Some of the typical models will be briefly summarized and described in Section II. A scoring model is just a single classifier, such as a decision tree, which has a fixed set of input speech features and is trained in a supervised learning way from the instances containing these speech features and human judged scores. However, this type of modeling approach has both theoretical and practical limitations in achieving accurate and robust score ratings.

Many years of research on the human rating process shows ample evidence for a considerable degree of variability among human raters. Milanovic et al. [2] assembled a heterogeneous list of elements human raters focused on when scoring written essays, such as the length of writing, accuracy of grammar, richness of vocabulary, etc. They found that the weight attributed to any particular element varied widely among raters during the rating process. Eckes [3] classified a group of human essay raters in a essay rating task into several distinct groups according to the features they used for scoring. He obtained several groups of raters and noted that each group had its own scoring profile (which focused on specific elements of writing during scoring). Therefore, given the fact that human raters may focus on varied writing/speaking elements, using a fixed set of features when building an automated scoring model may bring mismatches to human scoring results.

In addition, among the features extracted for use in building scoring models, some are highly predictive of human scores while others are "weak". "A few highly indicative features can swamp the contribution of many individually weaker features, even if the weaker features, taken together, are just as indicative of the output. Such a model is less robust, for the few strong features may be noisy or missing in the test data [4]." This issue may impact automated speech assessment since there is a gap between obtaining training data and applying the trained scoring model on real test data. In general, the data used to train a scoring model is from either pilot studies (or pretests) previous to the operational test or old test data obtained several months prior. Therefore, one issue to be expected is that the data used to build scoring models will be somewhat different to the data from the operational test. For example, test questions may be updated between the pilot studies and the operational test. Therefore, making it important to build a robust scoring model that can handle such variations.

In this paper, we will describe research to cope with the impact of scoring accuracy and robustness in automated speech assessment systems by applying a feature-bagging learning approach. The paper is organized as follows: Section II reviews the previous research on the application of a variety of classifiers to automatically score speech data and on using feature subsets to improve discriminative learning; Section III describes the basic automated speech scoring system and two approaches to build scoring models (using a single model vs. using feature-bagging); Section IV reports on our experiment

results on the three research questions; and finally, Section V discusses the experimental results and plans for future research works.

## II. Previous research

Several types of classifiers have been used to build scoring models for automated speech assessment or pronunciation verification systems. For example, SRI's EduSpeak system [5] used a decision-tree model to automatically produce a score from a set of discrete score labels. Pearson's Versant speaking test [6] used a non-liner model but did not disclose the details of their model. Educational Testing Service (ETS) has been working in the speaking assessment area and used both Support Vector Machine (SVM) and Classification and Regression Tree (CART) models in their initial trial [7] on scoring spontaneous speech. Recently, they have been using a Multiple Regression (MR) model in their systems given its simplicity in implementation and in explaining the rating process to the public [8]. In addition, other learning models, such as the Gaussian Mixture Model (GMM) [9] and the Linear Discrimination Analysis (LDA) [10] have been used in others' past research of pronunciation evaluation. From this brief summary, we can find that using a single classifier as the scoring model is still dominant in the automated speaking assessment research field.

In contrast, there has also been a trend of adopting ensemble learning in the machine learning research field to cope with classifier variances to further improve the classification accuracy and robustness. Ho [11] used an ensemble of decision trees trained on different feature subsets. Breiman [12] proposed the random forest method to use random decision trees by using a random feature at each decision tree building node. O'Sullivan et al. [13] proposed the FeatureBoost algorithm, which is analogous to AdaBoost. Their method dynamically adjusts the weights on features instead of on instances in order to fully utilize all available features. They showed that FeatureBoost is more robust than AdaBoost on synthetically corrupted UCI datasets. Bryll et al. [14] presented an attribute bagging (AB) technique for improving the accuracy and stability of induced classifier ensembles using random subsets of features. On a hand-pose recognition dataset, they compared the proposed AB method with conventional ensemble learning methods, such as bagging. Their results showed that AB gives consistently better results than bagging, both in accuracy and stability. Sutton et al. [4] introduced the concept of "underfitting" (see the quotation in Section I). To better utilize the rich features from NLP tasks, they applied the feature bagging in sequence-labeling tasks to prevent the issue of undertraining the weights. Compared to using a single model, their feature-bagging learning method showed improved accuracy and robustness.

Given the issues that impact the scoring model in automated speaking assessment systems and the success of using feature bagging on many different machine learning tasks, we expect that applying the feature-bagging approach can help to improve the accuracy and robustness of the scoring model.

## III. Methods

### A. Speaking Tests

TOEFL is a large-scale English test for assessing test-takers' ability to use English to study in colleges, using English as its primary teaching language. The TOEFL dataset was collected during operational testing. All responses were produced by real test takers and recorded in test centers in which certified standard equipments such as microphones, were used according to the test-delivery protocol. In each test session, test-takers were required to respond to six speaking test items, in which they were required to provide information or opinions on familiar topics, based on their personal experience or background knowledge. For example, the test-takers were asked to describe their opinions about living on- or off-campus.

To facilitate students' preparation for the TOEFL test, an online practice test, TOEFL Practice, was set up. Audio responses from the practice test (which used retired testing material from the TOEFL test) were also collected. However, some noticeable differences appeared between the two datasets. For example, test-takers for the TOEFL Practice took the practice test at home and used their own computers and audio input devices.

Each spoken response was assigned a score in the range of $1$ to $4$, or $0$ if the candidate either made no attempt to answer the item or produces a few words totally unrelated to the topic. Each spoken response could also be labeled as a "technical difficulty" (TD) when technical issues may have degraded the audio quality so that a fair evaluation was not possible. Note that in the experiments reported in this paper, we excluded both $0$ and TD responses from our analyses. In both datasets, the human scoring process used scoring rules in accordance with the scoring of the operational TOEFL test.

Table I describes the size and distribution of item scores for each dataset mentioned above. We can find that the scores were more evenly distributed on the TOEFL data than on the TOEFL Practice dataset. We expected that test-takers who had a sufficiently high skill level (the speaking score was more than $1$) considered using the practice test.

| Set | N | $N_{SC1}$ | $N_{SC2}$ | $N_{SC3}$ | $N_{SC4}$ |
|---|---|---|---|---|---|
| TOEFL | 5651 | 458 | 2098 | 2384 | 711 |
| TOEFL Practice | 1777 | 76 | 564 | 892 | 245 |

TABLE I
Human score distribution statistic of the TOEFL and TOEFL Practice test datasets

### B. Speech Recognition

To automatically score spontaneous speech, we used the method proposed in [15]. In this case, a speech recognizer is used to recognize non-native speech whereby a forced alignment is conducted based on the obtained recognition hypotheses. From recognition and alignment outputs, a number of features are extracted from multiple aspects, such as the timing profiles, recognition confidence scores, alignment likelihoods, an so on.

For speech recognition and forced alignment, we used a gender-independent fully continuous Hidden Markov Model (HMM) speech recognizer. Two different Acoustic Models (AM) were used in the recognition and forced alignment steps respectively. The AM used in the recognition was trained on approximately 30 hours of non-native speech from the TOEFL Practice test. For language model training, a large corpus of non-native speech (approximately 100 hours) was used and mixed with a large general-domain language model (trained from the Broadcast News (BN) corpus [16] of the Linguistic Data Consortium (LDC)). The AM used in the forced alignment was trained on native speech and high-scoring non-native speech. It was trained as follows: starting from a generic recognizer, which was trained on a large and varied native speech corpus, we adapted the AM using a batch-mode MAP adaptation. The adaptation corpus contained approximately $2,000$ responses with high scores in previous TOEFL Practice tests and the data from native English speakers who participated in a study related to the TOEFL test.

### C. Assessment features extraction

A construct is a set of knowledge, skills, and abilities that are measured by a test. The construct of the speaking test is embodied in the rubrics that human raters use to score the test. The rubrics used in the TOEFL generally consist of three key categories: *delivery*, *language use*, and *topic development*. Language use refers to the range, complexity, and precision of vocabulary and grammar use. Topic development refers to the coherence and fullness of the response. In practice, given the challenges in achieving adequate recognition accuracy on non-native spontaneous speech inputs, most of ASR-based speech assessment systems focus on the delivery aspect. The delivery in turn can be measured on four dimensions: fluency, intonation, rhythm, and pronunciation.

We extracted the following two types of features, including (1) speech features based on the speech recognition output as described in [8] and (2) pronunciation features that indicated the quality of phonemes and phoneme durations [15]. Among all extracted features, we selected 11 features that were found to be useful to indicate speaking proficiency levels. Table II lists the names, dimensions, and categories in the assessment construct, as well as the descriptions of these features. Details of computing these features can be found in [8], [15].

Table III reports Pearson correlation coefficients $r$ between each feature and human-judged scores on the two datasets used in this paper. We can find that some features, such as $L_6$ shows a quite high correlation (more than $0.5$) while some features have much lower correlations (such as *wdpchk* and *longpmn*). In addition, in general, most of features' $r$s are higher on the TOEFL dataset than on the TOEFL Practice dataset.

### D. Scoring models

We treated all scoring as a classification task, i.e., to learn a statistical model via a supervised learning method from a training dataset containing an input of speech features and an output of human-judged scores. Then, for each incoming spoken response, the learned model was applied on the features extracted from the audio file to obtain a score from 1 to 4. In our experiment, we used CART to create the classifiers. We compared the two approaches to build CART-based scoring models.

The first approach used a single CART tree. In contrast, for the second approach, we applied the feature bagging (or attribute bagging) method. The algorithm is shown as follows:

**Algorithm 1** the feature bagging ensemble learning using CART trees

Given a training data set, $Y = X, C$, where $Y$ represents the training instances containing features $X$ and scores $C$.
**for** $i = 1 \rightarrow T$ **do**
  Randomly selecting $m$ features from $M$ available features to generate a new training set, $Y_i$, with the reduced features
  Training a CART tree $H_i$ using $Y_i$
  $i \leftarrow i + 1$
**end for**
Using all $T$ CART trees ($H_1$ to $H_T$) to classify a new instance and outputting the final prediction by majority voting among $T$ outputs.

## IV. Experiments

### A. Setup

In this paper, the following three questions need to be answered regarding the feature-bagging learning method:

- does its application improve the scoring performance?
- does its application help to cope with the variaions of features we face in the testing stage?
- does its application help to apply the automated assessment system to new test data?

When comparing the two learning methods, single-CART vs. feature-bagging, we repeated 20 iterations of model building and testing on a given dataset. In our experiments, we used two metrics which have been widely used to evaluate the performance of automated scoring systems, including the

| | TOEFL Practice | TOEFL |
|---|---|---|
| wdpchk | 0.15 | 0.19 |
| wpsec | 0.44 | 0.56 |
| silpwd | -0.23 | -0.20 |
| silmean | -0.24 | -0.25 |
| longpmn | -0.13 | -0.12 |
| longpfreq | -0.28 | -0.30 |
| tpsec | 0.26 | 0.49 |
| tpsecutt | 0.38 | 0.56 |
| lmscore | 0.30 | 0.27 |
| $L_6$ | 0.51 | 0.60 |
| vowel_shift | 0.28 | 0.39 |

TABLE III
The Pearson correlation coefficients ($r$s) between the features and human scores on two datasets

| feature | dimension | category | description |
|---|---|---|---|
| *wdpchk* | fluency | delivery | word per chunk (speaking rate) |
| *wpsec* | fluency | delivery | word per second (speaking rate) |
| *silpwd* | fluency | delivery | silence per word |
| *silmean* | fluency | delivery | mean of silence durations |
| *longpmn* | fluency | delivery | mean of long pauses durations |
| *longpfreq* | fluency | delivery | long pause frequency |
| *tpsec* | fluency & vocabulary diversity | delivery & language use | unique words normalized by total word duration |
| *tpsecutt* | fluency & vocabulary diversity | delivery & language use | unique words normalized by utterance durations |
| *lmscore* | grammatical accuracy | language use | language model score |
| $L_6$ | pronunciation | delivery | average likelihood per second normalized by the rate of speech |
| $\bar{Sn}$ | pronunciation | delivery | average normalized vowel duration shifts |

TABLE II
A LIST OF SPEECH FEATURES USED IN OUR EXPERIMENTS

exact-agreement between the scores predicted by the scoring model and the scores judged by human raters as well as the Pearson correlation coefficient ($r$) between these two sets of scores. Among the measurement results from these 20 iterations, we used a $t$-test with a $p$ value of $0.05$ to decide which method is significantly better statistically.

For the first research question, using the TOEFL data, we ran 20 iterations of model building/evaluations using both learning methods. In each iteration, using stratified sampling based on a human-score distribution, $60\%$ of the instances were used for training and the remaining $40\%$ of the instances were used for evalution. For the second research question, we added noise on the instances in our testing set. Since the $L_6$ feature is highly indicative of predicting human scores (with a Pearson $r$ about 0.5), for each instance in the testing set, we randomly added noise ranging from $-SD(L_6)$ to $SD(L_6)$. Similar to our experiment in the first research question, we conducted model building/evaluation on 20 iterations of the TOEFL dataset. For the third research question, we trained the scoring models using the two learning methods from the TOEFL dataset and then applied the models on the other dataset (TOEFL Practice). From our brief description of the differences between the operational test and the practice test in Section III-A, we know that these two tests have different participants, data collection infrastructures, and varied score distributions. Similar to the two experiments above, we repeated 20 iterations of the model building/evaluations. For each iteration, we randomly selected $40\%$ of the instances from the TOEFL data and tested on the entire TOEFL Practice data set.

In our experiments, we used the J48 CART tree, an implementation of C4.5 tree in Java, in the Weka [17] machine learning toolkit. We only enabled the pruning option to further improve the classification accuracy but used all of the remaining default parameters. The feature bagging was implemented in R statistical software using the RWeka package [18]. We selected $T$ to be 20 and $m$ to be 5 based on our initial pilot study. The $t$-test statistics were also conducted in R.

### B. Experimental Results

Table IV reports the results of the three experiments described above. For the first research question (exp-1), we found that compared to the single-CART method, the application of the feature-bagging method increased the averaged exact-agreement from $54.69\%$ to $56.71\%$ and the Pearson $r$ from $0.526$ to $0.546$. Both of these performance improvements are statistically significant. A comparison of agreement and correlation results among 20 iterations for the exp-1 can be found in Figure 1, which shows the boxplots of exact-agreement and Pearson $r$ conditioned on these two learning methods. The feature-bagging method clearly shows a superior performance in the figure.

| Exp. | $\bar{Agr.}_{CART}$ | $\bar{Agr.}_{FB}$ | $t$-test $p$ |
|---|---|---|---|
| exp-1 | 54.69 | **56.71** | $p < 0.001$ |
| exp-2 | 53.59 | **56.98** | $p < 0.001$ |
| exp-3 | 49.77 | **52.20** | $p < 0.001$ |
| Exp. | $\bar{Corr.}_{CART}$ | $\bar{Corr.}_{FB}$ | $t$-test $p$ |
| exp-1 | 0.526 | **0.546** | $p = 0.003$ |
| exp-2 | 0.482 | **0.529** | $p < 0.001$ |
| exp-3 | 0.384 | **0.401** | $p < 0.038$ |

TABLE IV
EXPERIMENTAL RESULTS COMPARING THE SINGLE-CART (DENOTED AS CART) AND THE FEATURE-BAGGING (DENOTED AS FB) ON THE THREE RESEARCH QUESTIONS. BOTH EXACT-AGREEMENT (AGR.) AND PEARSON CORRELATION $r$ (CORR.) WERE REPORTED AS WELL AS $p$ VALUE FROM THE $t-test$ BETWEEN THESE TWO LEARNING METHODS. BOLD FONTS INDICATE THAT THE EVALUATION METRIC WAS STATISTICALLY HIGHER THAN THE ONE FROM USING SINGLE CART MODEL.

For the second research question (exp-2), when the most useful feature, $L_6$, was smeared during testing, we found that the feature-bagging method was impacted less than the single-CART method. Since an ensemble of CART trees with heterogeneous feature sets was used, for some CART trees in the ensemble, $L_6$ was excluded in the model training so that other weak features could have the opportunity to be used. As a result, when the dominant feature was smeared, the feature-bagging learning approach still worked by relying on other weak features. In addition, the results from the third experiment (exp-3) showed that the application of the feature-bagging method helped when using the trained model on a new test dataset. Consistent with the findings for the exp-2, using an ensemble of classifiers with heterogeneous feature sets allows all features, strong or weak, to be fully trained. Therefore, when some features drift on a new test dataset, the ability of using all the features help to keep scoring performance.

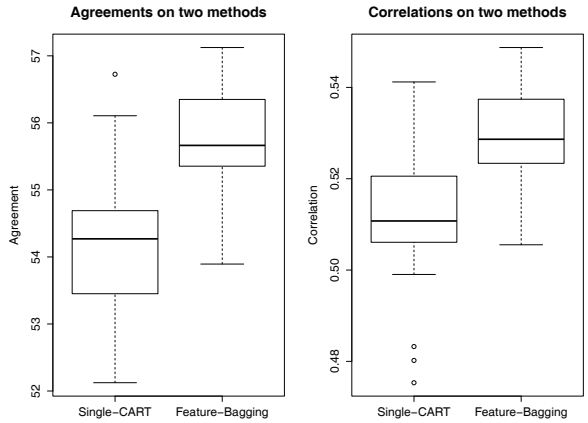


Fig. 1. Boxplots showing agreement and correlation $r$ obtained on two learning methods.

## V. Discussions

An increasing amount of research has been using ASR technology to build automated-speaking assessment systems. In these systems, the scoring model is a critical part, which converts the extracted speech features reflecting many aspects of speaking skills to scores, which are expected to be consistent and close to human-judged scores. However, for the scoring model built in a supervised learning model, some issues impact the model's performance, including the fact that human raters use heterogeneous features in their rating tasks and the variations across the datasets used in model training and testing. Although ensemble learning approaches (including bagging and boosting on instances and features) have been widely used in machine learning and ample evidence has shown improved performance in terms of accuracy and robustness, this new model-building methodology has not been adopted in the automated-speaking assessment research domain. To cope with the issues impacting the rating of human speech by machines, in this paper we conducted the first research applying feature-bagging to automated-speaking assessment research.

From our experiments, we clearly demonstrated the usefulness of the feature-bagging ensemble learning method. Compared to the traditional single-CART method, the feature-bagging method demonstrates statistically significant improved performance (measured by exact-agreement and Pearson $r$) from operational test data, TOEFL). We have also shown that the feature-bagging method can cope with feature variations that occur in the the test data (including noises and feature changes applied to new tests). Another attraction of feature-bagging is that it is easy to implement and embed in the current speaking assessment systems. This method can be used as a wrapper on the existing scoring models.

In future work, we will investigate how to optimize the process of generating an ensemble of classifiers. For example, we will use some feature-selection methods to optimize individual classifiers. In addition, instead of using the majority voting method to aggregate predictions from individual classifiers, we will investigate more sophisticated approaches such as considering the confidence of predicting each individual model.

## References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] M. Milanovic, N. Saville, and S. Shuhong, "A study of the decision-making behaviour of composition markers," *Studies in language testing*, vol. 3, p. 92–111, 1996.

[3] T. Eckes, "Rater types in writing performance assessments: a classification approach to rater variability," *Language Testing*, vol. 25, no. 2, p. 155, 2008.

[4] C. Sutton, M. Sindelar, and A. McCallum, "Reducing weight undertraining in structured discriminative learning," in *Proc. of HLT NAACL*, 2006, pp. 89–95.

[5] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications," *Language Testing*, vol. 27, no. 3, p. 401, 2010.

[6] J. Bernstein, A. V. Moere, and J. Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, p. 355, 2010.

[7] K. Zechner and I. Bejar, "Towards Automatic Scoring of Non-Native Spontaneous Speech," in *NAACL-HLT*, NewYork NY, 2006.

[8] X. Xi, D. Higgins, K. Zechner, and D. Williamson, "Automated Scoring of Spontaneous Speech Using SpeechRater v1.0," Educational Testing Service, Tech. Rep., 2008.

[9] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech and Language*, vol. 21, no. 6, pp. 219–230, 2007.

[10] C. Hacker, T. Cincarek, R. Grubn, S. Steidl, E. Noth, and H. Niemann, "Pronunciation Feature Extraction," in *Proceedings of DAGM 2005*, 2005.

[11] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, p. 832–844, 1998.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, p. 5–32, 2001.

[13] J. O'Sullivan, J. Langford, R. Caruana, and A. Blum, "Featureboost: A meta-learning algorithm that improves model robustness," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, p. 703–710.

[14] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, no. 6, p. 1291–1302, 2003.

[15] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[16] D. Graff, Z. Wu, R. MacIntyre, and M. Liberman, "The 1996 broadcast news speech and language-model corpus," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 11–14.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 1018, 2009.

[18] K. Hornik, C. Buchta, and A. Zeileis, "Open-source machine learning: R meets weka," *Computational Statistics*, vol. 24, no. 2, p. 225–232, 2009.