A Hybrid Approach to Automated Rating of Foreign Language Proficiency Using Oral Test Responses

Homayoon Beigi[1]
Recognition Technologies, Inc.
3616 Edgehill Road
Yorktown Heights, NY 10598
USA
beigi@RecognitionTechnologies.com

Abstract

This study was conducted to improve the automatic rating of oral test responses collected through Language Testing International's (LTI) Oral Proficiency Interviews using a Computer (OPIc). In OPIc tests, a computer automatically asks questions from the candidate and the responses of the candidate are recorded and consequently rated. This study has been performed on English OPIc tests. Although, no specific knowledge of the English language has been used for this phase of research and the results may be readily extended to tests in other languages. Preliminary results are quite promising, considering the utilization of the crude *Verbosity* feature.

Keywords: Language Proficiency Rating, Language Model, Oral Proficiency Interview, OPIc

1 Introduction

This study was conducted to improve the automatic rating of oral test responses collected through Language Testing International's (LTI) Oral Proficiency Interviews using a Computer (OPIc)[2]. This research is a continuation of the preliminary work reported at the end of 2008 in a technical report [2] by Recognition Technologies, Inc. In an OPIc test, a computer automatically asks questions from the candidate and the responses of the candidate are recorded and consequently rated. The results, reported here, stem from the application of the proposed rating technique on OPIc tests performed in English. However, no specific knowledge of the English language has been used in the algorithms presented for this phase of research. Therefore, the results may be readily extended to tests in other languages. Preliminary results are quite promising, considering the utilization of the crude *Verbosity* feature [2]. *Verbosity* is a function of the quantity of the speech which has been uttered in response to a test question. It uses no information about the content of the response.

The main objective of this research is to best mimic the rating style of human raters using an automated process. One of the goals of the project is to be able to increase the granularity of the ratings. Specifically, it is desired to be able to break down the IM rating to three subcategories (IM Low, IM Mid and IM High). The IM (Intermediate-Mid) rating encompasses a large portion of the population being tested. ACTFL defines these rating levels regularly in a guideline it provides its members [1]. To be able to attain this goal, the current rating style is learned by a statistical algorithm. Due to the continuous nature of the ratings returned by the statistical model, based on the *a-posteriori* probabilities returned by the model, one may increase the granularity of the rating to finer increments. This will produce further granularity which may not be reproduced by human raters in great ease!

Future phases of this research will be dealing with more substantial and qualitative features which utilize knowledge of the content of speech being uttered in response to the test questions.[2] Some more discussion will be provided at the conclusion of this chapter.

[1] Homayoon Beigi is the President of Recognition Technologies, Inc. and an Adjunct Professor of Computer Science and Mechanical Engineering at Columbia University

[2] These tests have been designed by the American Council on the Teaching of Foreign Languages (ACTFL) [1]

2 OPIc

An OPIc test does not have any human tester associated with it. One may say that the computer is the tester. First, the candidate makes a self-assessment of his/her language proficiency. Then, a test is automatically created for the candidate by combining a collection of Novice, Intermediate and Advanced prompts which are played back and the candidate is expected to respond to them. These responses are recorded and used to rate the candidate's proficiency. Different prompt categories require different lengths of response. The candidate is only allowed to produce a response which is limited in its length by some number of seconds, dictated by the test designers, corresponding to each prompt category.

Prompt Category	Level	Maximum Duration of Response (s)	Remark
nov1	Novice	30	All Novice level questions.
intrp	Intermediate	90	Role Play.
int2	Intermediate	60	Describe an object or a place.
int2n	Intermediate	60	A simpler version of int2 questions.
int3	Intermediate	60	Describe a process.
int3q	Intermediate	60	Intermediate Prompt to ask a question related to the intrp role-play.
adv1	Advanced	120	A past description.
adv2	Advanced	120	A past narration.
adv3	Advanced	120	Complication following the intrp role play.
adv4	Advanced	120	Description and narration following a story.
adv5	Advanced	120	Past description beyond the person (such as developments and changes).
adv6	Advanced	120	Past description beyond the person (such as a current event).

Table 1: Rating Levels in an OPIc Exam

Each OPIc exam consists of roughly 14 queries (prompts) that are picked from a large collection of stock questions. The questions are categorized into different levels of difficulty as well as the mental tasks that are required of the candidate in his/her response. Table 1 presents these categories as well as the maximum number of seconds allowed for the corresponding response and a quick remark about each category.

Depending on the self-proclaimed proficiency level of the candidate, a test is generated by combining a random set of questions, coming from the categories listed in Table 1. The number of questions from each category is dictated by the test level. For example, an Advanced test will have more advanced prompts in it, but it also includes some intermediate prompts. A Novice test will only have questions from the Novice and Intermediate categories. An Intermediate test will have more Intermediate questions (prompts) than an Advanced test would, but it will include less Advanced level prompts.

2.1 Audio Quality

The audio data was recorded using the μ-Law amplitude coding technique [5] at a sampling rate of 8 kilo Hertz (kHz). The audio was then immediately converted to the High Efficiency-AAC Audio Format (**HE-AAC**) which is a very aggressive, lossy and low-bit-rate audio compression technique.[3] The compressed audio was uploaded to a server. In a limited number of tests, the audio was converted into MPEG-1 Audio Layer 3 (MP3) instead of HE-AAC. All audio responses, in turn, were converted back to Mu-Law 8-kHz audio and subsequently converted to a 16-bit linear Pulse Code Modulation (**PCM**) form which was used in the recognizer for obtaining the features described here.

2.2 The Rating Process

OPIc tests are manually graded by human raters. There are 7 possible rating levels in the manual process. Table 2 shows the acronyms and numerical values used for the different rating levels. 1 corresponds to the least proficient group of speakers and 7 is associated with the highest level of proficiency.

2.3 Computed Features

Since the candidate responds to predefined prompts, his/her audio is not multiplexed with any other audio and is separately available for each response. Therefore, the *Verbosity* is computed by using the RecoMadeEasy® engine of Recognition Technologies, Inc. to extract segments where audio is present. The length in number of seconds of spoken audio constitutes

Proficiency Level	Acronym	Rating Level
Novice Low	NL	1
Novice Mid	NM	2
Novice High	NH	3
Intermediate Low	IL	4
Intermediate Mid	IM	5
Intermediate High	IH	6
Advanced	A	7

Table 2: Rating Levels in an OPIc Exam

Verbosity. To account for the length of each response (including pauses), the Verbosity feature is represented as a two dimensional vector which includes the length of spoken audio in seconds as the first dimension and the total length of the audio segment as the second dimension of the feature vector.

A rating process was trained and tested using the *Verbosity* feature. Let us assume that the feature for the t^{th} response is denoted by $\mathbf{f}_t : 1 \mapsto \mathscr{R}^2$ and that the prompt associated with that response is denoted by l_i. Also, let r_k denote the k^{th} rating level presented in Table 2. Theoretically, it is possible to describe any complex distribution by an infinite number of Gaussian distributions. However, in practice, this number may be made finite while obtaining a good approximation to the original complex distribution. If we assume that there exist a certain number of Gaussian Prototypes, the mixture of which describes the distribution of the features associated with responses to the i^{th} prompt category, then the *a-posteriori* probability of the rating given a feature $\mathbf{f}_t$ computed jointly for prompt l_i may be estimated by equation 1.

$$p(r_k|\mathbf{f}_t, l_i) = \sum_{j=1}^{N_i} p(r_k|g_j^i)P(g_j^i|\mathbf{f}_t, l_i) \tag{1}$$

Where g_j^i is the j^{th} Gaussian prototype for the i^{th} prompt category and N_i is the number of Gaussian prototypes used to map the features associated with prompt level l_i. The *a-posteriori* probability of the Gaussian Cluster, g_i given the feature $\mathbf{f}_i$ is given by equation 2.[4]

$$P(g_j^i|\mathbf{f}_t, l_i) = \frac{p(\mathbf{f}_t|g_j^i)P(g_j^i)}{p(\mathbf{f}_t, l_i)} \tag{2}$$

Where, $P(g_j^i)\ \forall\ j = 1, 2, \ldots, N_i$ is the set of *a-priori* probabilities (priors) estimated by a clustering technique. $p(\mathbf{f}_t)\ \forall\ t = 1, 2, \ldots$ are assumed to be 1 since at any instance, t, this represents the probability of occurrence of feature $\mathbf{f}_t$. Since the probability of occurrence of the feature vector, $\mathbf{f}_t$, has no bearing on the choice rating, r_k, this probability may be set to 1 at the time of the test. In other words, it is only important that,

$$P(g_j^i|\mathbf{f}_t, l_i) \propto p(\mathbf{f}_t|g_j^i)P(g_j^i) \tag{3}$$

$p(\mathbf{f}_t, l_i|g_j^i)$ is the joint likelihood of $\mathbf{f}_t$ for prompt level l_i given Gaussian prototype g_j^i and may be computed using the equation for the *Normal* distribution,

$$\begin{aligned} p(\mathbf{f}_t, l_i|g_j^i) \quad = \quad & \frac{1}{(2\pi)^{\frac{d}{2}}\,|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \\ & exp\left\{-\frac{1}{2}(\mathbf{f}_t - \boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{f}_t - \boldsymbol{\mu}_i)\right\} \\ & where \begin{cases} \mathbf{f}_t,\ \boldsymbol{\mu}_i \in \mathscr{R}^d \\ \mathbf{\Sigma}_i : \mathscr{R}^d \mapsto \mathscr{R}^d \end{cases} \end{aligned} \tag{4}$$

In 4, $\boldsymbol{\mu}_i$ is the mean vector associated with the features of prompt level l_i, where,

$$\boldsymbol{\mu}_i \stackrel{\Delta}{=} \mathscr{E}\{\mathbf{f}_t\} \stackrel{\Delta}{=} \int_{-\infty}^{\infty} \mathbf{f}_t \; p(\mathbf{f}_t) d\mathbf{f}_t \tag{5}$$

The variance matrix of a multi-dimensional random variable is defined as,

$$\begin{aligned} \boldsymbol{\Sigma}_i &\stackrel{\Delta}{=} \mathscr{E}\left\{(\mathbf{f}_t - \mathscr{E}\{\mathbf{f}_t\})\,(\mathbf{f}_t - \mathscr{E}\{\mathbf{f}_t\})^T\right\} &(6)\\ &= \mathscr{E}\left\{\mathbf{f}_t {\mathbf{f}_t}^T\right\} - \boldsymbol{\mu}_i {\boldsymbol{\mu}_i}^T &(7)\end{aligned}$$

2.4 Training Process

In Equation 1, the mixture coefficients, $p(r_k|g_j^i)$ are computed at the training stage using the *Expectation Maximization* algorithm and in conjunction with the training data. Let us assume the total number of feature vectors associated with the rating, r_k, to be represented by T_{r_k}. Then, the joint likelihood of the feature vector, $\mathbf{f}_t$, associated with the training label, r_k, may be defined as,

$$\mathscr{L}_{t,k,j,i} = p(\mathbf{f}_t, r_k|g_j^i) \tag{8}$$

Therefore, the conditional probability of rating r_k with respect to the Gaussian prototype g_j^i, associated with prompt category, l_i is given by Equation 9.

$$p(r_k|g_j^i) = \frac{\frac{\sum_{t,k,i} \mathscr{L}_{t,k,j,i}}{\sum_{j=1}^{N_i} \mathscr{L}_{t,k,j,i}}}{T_{r_k}} \tag{9}$$

The *a-posteriori* probability, $p(r_k|\mathbf{f}_t, l_i)$ is given by Equations 3 and 9, with the prior probabilities also computed at the training stage.

Consequently, the posterior likelihood for any rating given the selected feature (Verbosity) may be computed. These values will not generally add up to one and are considered to be likelihoods. A normalization is done to impose the summation of 1, rendering the computed values akin to a probability. Then, the rating with the highest likelihood is taken to be considered as the final rating for that response. An averaging or voting method may be used among the several responses in a test, to come up with the final rating for the test.

2.5 Results

Figure 1 shows the number of test responses used for each prompt category in obtaining the results. As it may be seen from the bar chart, there is a vast bias in the number of test responses available for each prompt category. This should somewhat affect the results. Figure 2 shows three graphs, summarizing the preliminary results obtained at the response level. Accuracy of the rating is reported separately for each prompt category. In total, 973,204 test responses were used in obtaining the rating results. Since the amount of data was limited, a round-robin approach for data-conservation was used in selecting the training and test data sets. In this approach, the responses were split into 10 groups with balanced memberships from the different prompt categories. Then, 10 tests were designed, in each of which, 9 segments were used for training and the disjoint remaining set was used for testing. A round-robin rotation allowed us to utilize the whole data for both training and testing while keeping the training and test data independent at all time.

The accuracy of reproducing the human ratings is shown to be slightly above 24% in average over all the 12 prompt levels of Table 1. Figure 2 also reveals two more statistics about the automatic rating performance. The first statistic is the graph labeled, "Accurate within 1". This graph shows the accuracy of the automatic rating when compared with the human rating, if one allowed the results to be off by 1 level within the possible 7 rating levels. On average, 66.33% of the ratings are correct when this looser requirement is adopted. For novice prompts this number is as high as 74.14% and it reaches its lowest performance for the most advanced prompts, hovering around 51%. The significance of this statistic becomes more clear when a test is rated both manually and through this algorithm. If the algorithm assigns a rating, higher by 1 level, compared to the human rating for candidate *A* and assigns a rating lower by 1 level, compared to the human rating given to candidate *B*, when the human ratings of *A* and *B* are the same, then, with about 66% confidence, candidate *A* is superior

to candidate *B*. This may be extended to include 2 levels of relaxation, as reported in Figure 2 with the graph labeled, "Accurate within 2". In this case, the average confidence is about 91% for over all the different prompt levels. It can be as high as 98% for *int*2*n* level prompts and no less than about 84% for *int*3*q* prompt. Table 3 shows the detailed percentages associated with the results of Figure 2.

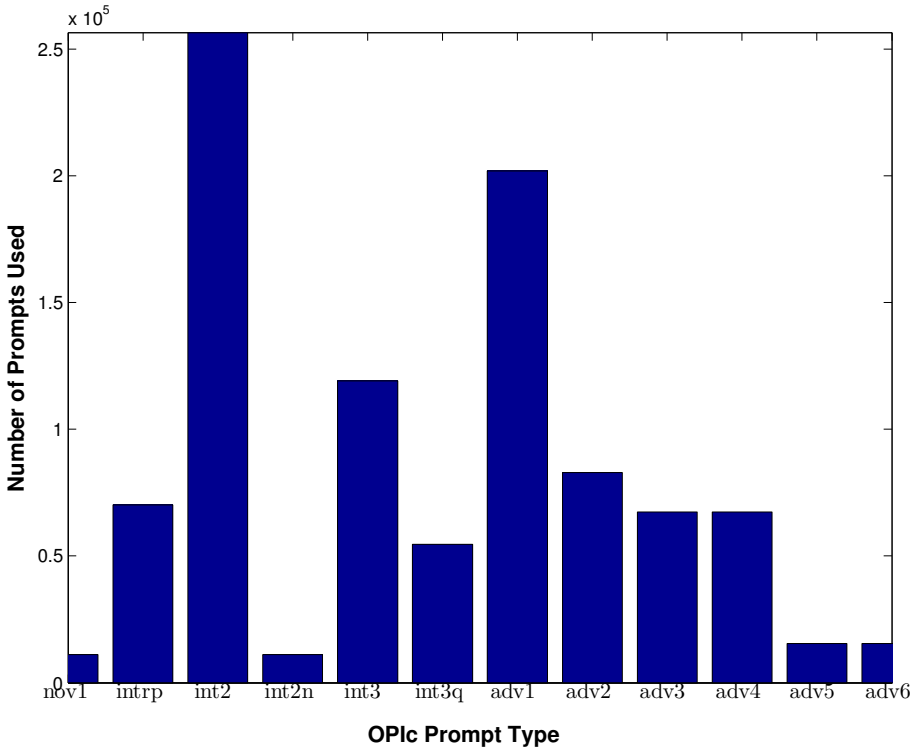

Figure 1: Number of test samples used for the different prompt categories in obtaining accuracy results

Prompt Category	Rating Accuracy	Accurate within 1	Accurate within 2
nov1	32.51%	74.14%	96.06%
intrp	26.20%	67.80%	90.62%
int2	23.67%	66.37%	91.15%
int2n	33.41%	80.23%	97.77%
int3	26.39%	71.00%	91.46%
int3q	21.88%	60.97%	83.80%
adv1	25.48%	69.59%	91.59%
adv2	24.69%	66.47%	90.67%
adv3	24.47%	66.24%	91.34%
adv4	25.06%	68.41%	91.46%
adv5	10.87%	50.89%	88.97%
adv6	13.40%	53.89%	85.24%

Table 3: OPIc automatic rating performance for the different prompt categories

3 Extension to the IM Delineation Analysis

In Section 2.5, the methodology for reproducing a rating at the same granularity as that of human raters was reported. In reality, the ratings returned by the proposed system are produced in the form of a distribution. Using the provided distribution, further rating resolution may be implemented. Figure 3 shows an example of the likelihood distribution returned by the automatic rating engine. As discussed in Section 2.5, depending on the deviation of the automatic rating from the manual rating, a finer resolution may be attained. For example, in the rating level of interest, namely *IM* (rating 5), all the manual results are, by definition, rated *IM*. This means that if the automatic system rates the candidate higher, then

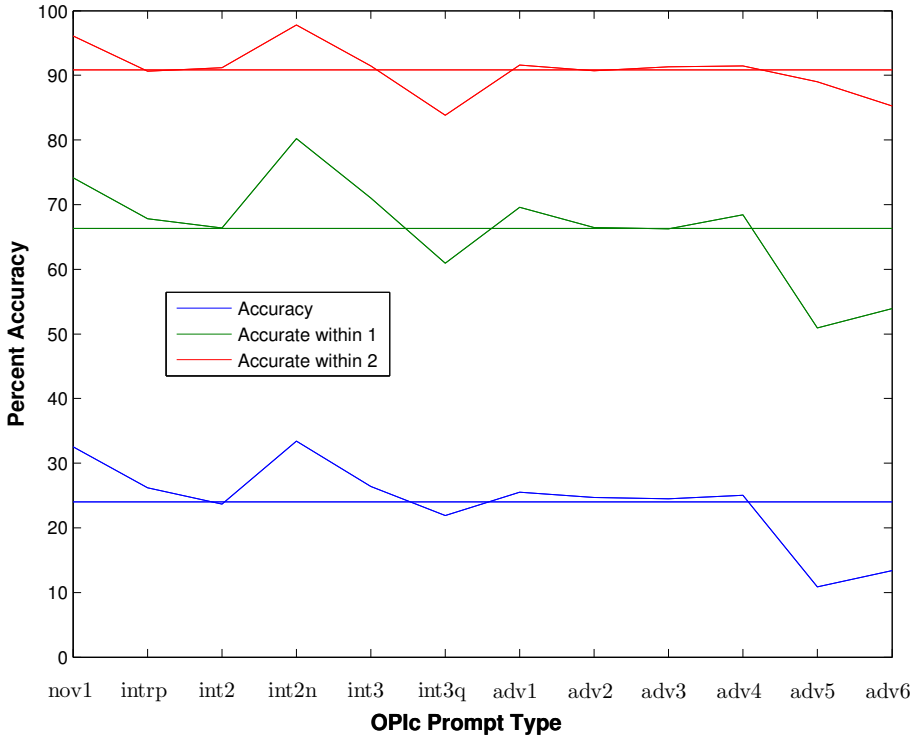

Figure 2: OPIc Test Accuracy using the Verbosity Feature

the candidate may be considered an *IM High* (delineation level 3) and if a lower rating is produced by the automatic system, then the candidate may be considered as an *IM Low* (delineation level 1). By looking at the probability levels returned by the automatic system, cutoff levels for delineation levels 1 and 3 may be obtained, thereby allowing for the determination of the range of level 2 delineation. The confidence in this rating is given by the last two columns of Table 3, depending on how much the automatic rating deviates from the manual rating. This translates to an average confidence of 66% if the discrepancy is at least one rating level and about 91% if the discrepancy is at least two levels.

The next section will be concerned with the development of techniques for combining the 14 or so responses for each test to produce an overall rating for the candidate. In the process, a methodology is presented for the evaluation of the delineations associated with *IM*-rated candidates. It utilizes the likelihood distributions depicted by Figure 3 in association with the manual rating (IM) to evaluate the delineation for the candidate. For instance, if the distribution leans more toward the higher rating, then an *IM High* (delineation level 3) is recommended. A confidence level is also presented to back the choice of recommended delineation.

4 Combination to achieve test-level rating

In this section, we will combine the automatic ratings from different responses in a test, to come up with a final test-level rating, r. Each test is made up of about 14 responses to questions which are chosen from the different *Prompt Categories* listed in table 3. In the previous sections, we discussed the methodology for computing the automatic rating of each response in the test, conditioned on the prompt category. The resulting rating is a likelihood distribution associated with the different possible ratings. To come up with a final rating for each test, the mean distribution is computed over all the individual prompt responses in the test. The resulting distribution is used to assign a rating to the test. Once the mean distribution is computed, the rating is the mean value of the mean distribution. This rating, r, is a real number and can vary from 1 (*novice*) to 7 (*advanced*). If an integer rating is desired, r would have to be rounded to the closest integer.

In order to evaluate the performance of the automatic rating, each test has been evaluated by a human rater as well. This rating is denoted by, r_H. Table 4 shows the accuracy of the automatic rating, r, when compared with the human rating, r_H. The first row of the table shows the accuracy for a rounded integer version of the rating which means that $|r - r_H| < 0.5$. The mean accuracy, seems to be over 53% which is considerably higher than the mean accuracy of about 22% associated with each prompt response.

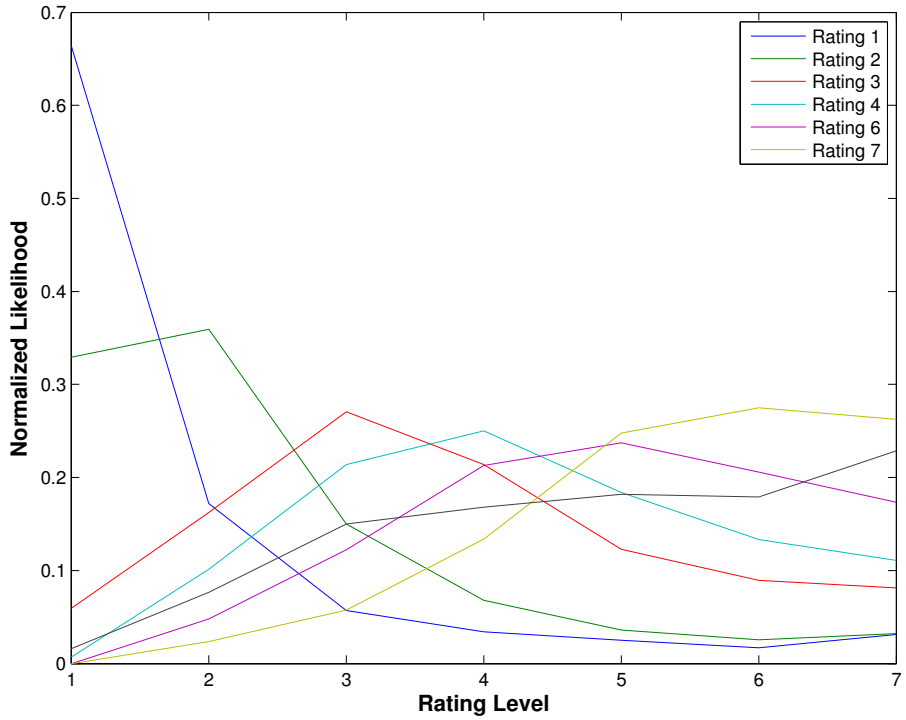

Figure 3: Likelihood distribution for the rating levels of test samples associated with prompt category *int*2 and ratings 1 through 7

Much higher accuracies are achieved if one loosens the performance criterion. For example, in *row* 2, of Table 4, about 86% accuracy is achieved if one allows the difference between the automatic rating and the human rating to be within one rating level.

Criterion	Accuracy
$\|r - r_H\| < 0.5$	53.21%
$\|r - r_H\| < 1.0$	85.81%
$\|r - r_H\| < 2.0$	98.60%
$\|r - r_H\| < 3.0$	99.51%

Table 4: Performance of the OPIc automatic combined (test-level) rating

A total of 5138 tests were rated using the automated system discussed in this report. Figure 4 shows the the histogram of the deviation of the automatic rating from human ratings. The mean deviation is $\mu = -0.056$ and the standard deviation is $\sigma = 0.74$. Note that a deviation of up to 0.5 is considered a correct rating when integer rating is used. This amounts to the 53.21% accuracy reported in row 1 of Table 4.

5 Conclusion

In the study covered in this reported, it has been shown that we are able to duplicate the human rating of oral proficiency tests through a computer using an automated process with very crude features. The features, studied here, are only related to the amount of speech generated by candidates while responding to questions. Even using this crude information, over 53% accuracy is achieved in duplicating human rating results. In addition, if one allows for an error of less than one level within the 7 possible rating levels, over 86% accuracy is achieved. The distribution of the deviation from the human rating, reported in Figure 4, shows a slight bias toward under-rating ($\mu = -0.056$). This is also apparent from the slight skew of the distribution toward the negative deviation. The existence of most of the tests within the middle three bins of the distribution represents the sanity of this rating. Namely, it shows that there are not many surprises in the automatic rating results. This may also be observed by the fact that about 99% of all the tests fall within a deviation of up to 2 rating levels when compared to human rating.

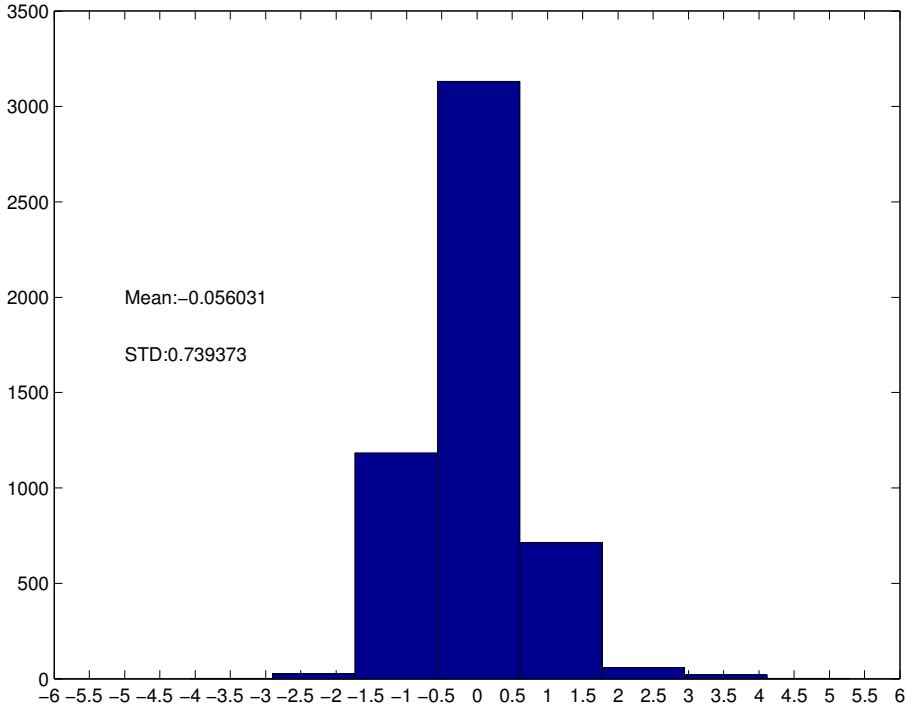

Figure 4: Deviation statistics of the automatic ratings versus human rating

One of the objectives of this study is to achieve better subrating of the Intermediate-Mid level tests. Since most candidates are categorized in this level, it is highly desirable to break up this category into smaller groups. In Section 3, it was noted that based on the combined results from the human rating and the automatic rating, a higher granularity may be achieved for further separation of the Intermediate-Mid level candidates into 3 subcategories. This is done by using the results from the automatic rating system to separate the candidates who have been human-rated into the Intermediate-Mid level. Since human raters are generally not able to rate candidates in a higher granularity than the 7 levels shown in Table 4, it is hard to assess the performance of the automatic subrating. However, based on the tight performance of the distribution (small standard deviation) shown in Figure 4, it may be argued in a qualitative fashion that very reliable results may be obtained. Such combination techniques have been used in many mathematical applications such as fusion results shown in [6].

These results are quite promising, since the basic accuracy for each prompt response was shown to be below 23% in the preliminary results, but using a combination of the results from the roughly 14 prompt responses per test, the reported respectable accuracies of over 53% have been achieved. Based on the preliminary studies reported in [2] and the results of this report, a considerably more accurate system is foreseen once linguistic information is utilized in the calculation of the automatic rating of the OPIc tests.

The author has been working on creating a language model based on ACTFL *Writing Proficiency Tests (WPT)* [1]. Also, work is in progress to produce the transcription of the discussed oral tests using a speaker independent speech recognizer. The transcribed output will then be processed by the said language model, in order to produce ratings based on linguistic content. The results will then be combined with those discussed here, in order to produce a hybrid system for automated rating of foreign language proficiency. The language model is being developed based on WPTs to avoid further complications due to transcription errors. Most speech recognizers will not be able to produce better than 50% word-level accuracy on speech of non-native speakers. Therefore, it is important to develop the language model independent of these transcription errors. Part of the practical problem at this moment is the number of WPTs which are available for developing this language model. In the absence of a large number of WPTs, massive amount of text from large literary corpora are being used to produce a target model. Although such target model would have to be modified to fit the statistics coming from non-native language production. In addition, since written language is governed by somewhat different statistics compared to oral language, in most languages, the produced language model would have to be modified to reflect oral speech.

The methods discussed here are mostly language-independent, although relevant statistics would need to be generated for the language of interest to achieve greater accuracy. The path of the future work, discussed in the above paragraph, is far more language-dependent and requires a speech recognizer and a pertinent language model in the language of choice. The performance of the language model and rating by textual context is plagued by many practical uncertainties produced by

the behavior of the speech recognizer and the inconsistencies among written and oral language as well as different language proficiency levels. In addition to statistical analysis on the textual context, it will definitely be necessary to devise methods for finding common errors among non-native speakers of the language of choice.

6 Acknowledgments

The author would like to thank the Center for Language Studies at the Brigham Young University, Language Testing International (LTI) and the American Council on the Teaching of Foreign Languages (ACTFL) for making this research possible.

References

[1] ACTFL: American Council on the Teaching of Foreign Languages (ACTFL) Guidelines (2012)

[2] Beigi, H.: Whether Computer Analyses Can Predict Human Ratings of Speaking Proficiency. Recognition Technologies, Inc. (2008). Technical Report No.: RTI-20081205-01

[3] Beigi, H.: Fundamentals of Speaker Recognition. Springer, New York (2011). ISBN: 978-0-387-77591-3, http://www.FundamentalsOfSpeakerRecognition.org

[4] Duda Richard O.; Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley and Sons, New York (1973). ISBN: 0-471-22361-1

[5] G.711: Pulse Code Modulation (PCM) of Voice Frequencies. ITU-T Recommendation (1988). URL www.itu.int/rec/T-REC-G.711/e

[6] Viswanathan, M., Beigi, H.S., Maali, F.: Information Access Using Speech, Speaker and Face Recognition. In: IEEE International Conference on Multimedia and Expo (ICME2000) (2000)