Whether Computer Analyses Can Predict Human Ratings of Speaking Proficiency

Recognition Technologies Inc.
Technical Report: RTI-20081205-01

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

Abstract

This study was conducted in order to answer the question, "Can Computer Analyses Predict Human Ratings of Speaking Proficiency." The study treats two different types of Oral Proficiency Interviews (OPI)[2], 1. a telephone-administered OPI where the conversation between a rater and a candidate is being recorded and later rated for proficiency and 2. a computer-administered test (OPIc) where the computer asks questions and the responses of the candidate are recorded and consequently rated. Two features based on the quantity and the quality of the responses of the candidate are introduced and then assessed for effectiveness. English has been selected as the language for which to do this preliminary study. The Speaker Recognition Engine of Recognition Technologies, Inc. and the Speech Recognizer of IBM T.J. Watson Research Center are used to compute the proposed features. Initial results are quite promising and show discrimination capabilities using these features of up to 91% on the training data.

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

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

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

1 Introduction

This study was conducted in order to answer the question, "Can Computer Analyses Predict Human Ratings of Speaking Proficiency?" The study treats two different types of Oral Proficiency Interviews (OPI), 1. a telephone-administered OPI, simply referred to as, "OPI," where the conversation between a rater and a candidate is recorded and later rated for proficiency and 2. a computer-administered test, "OPIc," where the computer asks questions and the responses of the candidate are recorded and consequently rated. English has been selected as the language for which to do this preliminary study. Two features based on the quantity and the quality of the responses of the candidate are introduced and then assessed for effectiveness.

The following two sections treat these two modes of testing (OPI and OPIc) and develop the features for automatic rating of the candidates' proficiency level based on the candidates' responses. In each section, after describing the test scenario, the theory and motivation behind the development of the features are discussed. Each section produces results for the experiments geared toward showing the effectiveness of the proposed features for conducting automatic rating.

Finally, concluding remarks are presented and the path for future work is laid. Also, future challenges and expectations are foreseen and discussed.

2 OPI

In an ACTFL OPI, the candidate calls an Interactive Voice Response (IVR) system and enters his/her access code which has been provided to him/her in advance. The IVR system uses the access code to query the database for information about the test. The IVR system will then call the tester and will connect the candidate and the tester while recording their conversation. The conversation is then made available to multiple raters who may either call into the IVR system or access the audio through the Internet-Based back-end of the system. The raters will, in turn, submit their rating for the candidate. The objective of this section is to assess the possibility of automating this rating process. Usually two raters grade each candidate; but if the two raters disagree, a third rater is used to resolve the disagreement. To be able to assess the automatic rating process, only tests were selected for which there was agreement between the first two raters.

An average OPI test lasts about 30 minutes. The way the audio has been recorded, the tester's and the candidate's audio have been multiplexed into a single channel. The first objective is to separate the audio into segments representing the tester and the candidate's contribution. The venue of these tests is such that usually, the tester reads out a one-minute informational excerpt including disclaimers. Then, the tester will start by asking questions to which the candidate will respond. The conversation will then take turn between the tester and the candidate until the end of the test.

In this study, we are concerned with three different ratings presented in Table 1. A total of 726 tests were selected which were rated by at least two raters and for which the all the raters agreed on the proficiency level of the candidate. These tests were conducted by a total of 27 testers. The candidates were all distinct in the set. The testers did not conduct a uniform number of tests. The maximum number of tests conducted by a tester was 168 and the minimum was 1 with mean 26.9 and standard deviation 37.2.

Proficiency Level	Rating
Intermediate	1
Advanced	2
Superior	3

Table 1: Ratings used in the OPI tests

$\mathscr{L}_t$	*Log Likelihood – Tester Identification*
$\mathscr{L}_c$	*Log Likelihood – Candidate Identification*
$\mathscr{C}$	*Identification Confidence*
$S_{t\,i}$	*i^{th} Tester Segment (Seconds)*
$S_{c\,i}$	*i^{th} Candidate Segment (Seconds)*
V_t	*Tester Verbosity (Seconds)*
V_c	*Candidate Verbosity (Seconds)*
V	*Total Verbosity (Seconds)*
N_t	*Number of Tester Segments*
v_t	*Mean Tester Verbosity (Seconds)*
σ_t	*STD of Tester Verbosity* (Seconds)
N_c	*Number of Customer Segments*
v_c	*Mean Candidate Verbosity (Seconds)*
σ_c	*STD of Candidate Verbosity* (Seconds)
v	*Mean Verbosity (Seconds)*
σ	*STD of Verbosity* (Seconds)
$\mathscr{V}_t$	*Tester Relative Verbosity* ($\frac{V_t}{V}$)
$\mathscr{V}_c$	*Candidate Relative Verbosity* ($\frac{V_c}{V}$)
υ_t	*Tester Relative Mean Verbosity* ($\frac{v_t}{v}$)
υ_c	*Candidate Relative Mean Verbosity* ($\frac{v_c}{v}$)

Table 2: Nomenclature related to Quantitative Features

2.1 Computed Features

Tables 2 and 3 show the nomenclature and define the values which were computed. In general, there are two major types of features which may be associated with Fluency, namely, Quantitative Features (Table 2) and Qualitative Features (Table 3). Quantitative features are those which evaluate the prolificness of the candidate's speech. The idea behind these features is to be able to evaluate how verbose the responses are and that candidates who can produce longer responses tend to be more proficient in the language. The term *Verbosity* has been used here to describe these quantitative features.

However, quantity is not the only measure of proficiency. It is also important that the candidate has a good command of the language and that he/she uses proper grammar and a colorful vocabulary. At this stage of our study, we are not considering any complex grammars. We have, however, considered the vocabulary skills of the candidate in features we call *Rareness*.

2.1.1 Quantitative Features – *Verbosity*

To compute the *Verbosity* of the candidate (V_c), the conversation had to be segmented into interlaced segments of the Tester's and the candidate's speech segments (S_{ti} and S_{tj} respectively). To achieve this goal, the Speaker Segmentation and Speaker Identification capabilities of the RecoMadeEasy$^{(TM)}$ engine of Recognition Technologies, Inc. were utilized.[1] First, Speaker Segmentation was used to segment the conversation into interlaced speech segments tagged with *Speaker A* and *Speaker B*. Next step was the identification of generic speakers *A* and *B* as the *tester* and the *candidate*.

To do this, the Speaker Identification capability of RecoMadeEasy$^{(TM)}$ was used in the following manner.[1] For each test, the tester is known from the database. A one-minute segment of audio from each tester was enrolled into the Speaker Recognition Engine (RecoMadeEasy$^{(TM)}$) and labeled with the initials of the tester. The enrollment generates a speaker model for each tester. Once the segmentation was completed, the pieces of audio from speakers A and B were presented to the Identification Engine and the closest match to the tester's speaker model was labeled with the tester's initials. Once either A or B had been labeled with the tester's initials, the other speaker was assumed to be the candidate and duly tagged.

The output of this process is in the form of multiple segments of audio identified by a beginning and an end for each segment in seconds of audio relative to the beginning of the conversation. Also, each segment was labeled by either the tester's initials or the word *Candidate*.

Please refer to Table 2. $\mathscr{L}_t$ and $\mathscr{L}_c$ denote the Log-Likelihoods of the audio segment belonging to the tester and the candidate respectively. The highest Log-Likelihood would indicate the label for the segment. Once the identification is performed, a confidence level is attached to the identification results given by equation 1. Since $\mathscr{L}_t$ and $\mathscr{L}_c$ are logs of likelihoods, the confidence level is proportional to the *log* of the likelihood ratio of the *Tester* to the *Candidate*.

$$\mathscr{C} = 100(\mathscr{L}_t - \mathscr{L}_c) \quad (1)$$

The *Verbosity* is given by equation 2.

$$\begin{aligned} V_t &= \sum_{i=1}^{N_t} S_{ti} \\ V_c &= \sum_{i=1}^{N_c} S_{ci} \\ V &= V_t + V_c = \sum_{i=1}^{N_t} S_{ti} + \sum_{i=1}^{N_c} S_{ci} \end{aligned} \quad (2)$$

Table 2 presents all the features related to *Verbosity* (V).

2.1.2 Qualitative Features – *Rareness and Complexity*

Table 3 shows a list of quality-related features. A Unigram for the English language was generated from a large corpus of over $290,000,000$ words. the Unigram basically shows the frequency of the words in this corpus. Over $270,000$ distinct words were extracted with their corresponding frequencies, $\mathscr{F}_w$. We define the *Rareness* feature for every word uttered by the candidate and the tester to be inversely proportional to $\mathscr{F}_w$. In fact we chose the constant of proportionality to be $1,000$, namely,

$$\mathscr{R}_w = \frac{1000}{\mathscr{F}_w} \quad (3)$$

For the qualitative analysis, due to limited resources, a subset of the OPI tests was chosen. 22 tests from each rating defined in Table 1 were randomly chosen to make up a pool of 66 tests. All the segments of audio recognized to be the candidate's audio by the method of the previous section were augmented together to make up the candidate's total speech. The same thing was done for the tester of each test. The audio clips were then submitted to a server-based copy of the $65,000$ vocabulary speaker-independent speech recognizer of the Human Language Technologies group of IBM Research, courtesy of that group. The resulting transcription was then processed using equation 3 to compute the rareness for each word in the transcribed text followed by the computation of *Complexity* using equation 4.

$$\begin{aligned} \mathscr{X}_t &= \sum_{w=1}^{\mathscr{W}_t} \mathscr{R}_w \\ \mathscr{X}_c &= \sum_{w=1}^{\mathscr{W}_c} \mathscr{R}_w \\ \mathscr{X} &= \sum_{w=1}^{\mathscr{W}_t+\mathscr{W}_c} \mathscr{R}_w \end{aligned} \tag{4}$$

where $\mathscr{W}_t$ and $\mathscr{W}_c$ are the number of words spoken by the tester and the candidate respectively.

Other related features were also computed. See the equations 5 and 6.

$$\begin{aligned} \dot{\mathscr{X}}_t &= \frac{\mathscr{X}_t}{V_t} \\ \bar{\mathscr{X}_t} &= \frac{\mathscr{X}_t}{W_t} \\ \dot{\mathscr{W}}_t &= \frac{\mathscr{W}_t}{V_t} \end{aligned} \tag{5}$$

$$\begin{aligned} \dot{\mathscr{X}}_c &= \frac{\mathscr{X}_c}{V_c} \\ \bar{\mathscr{X}_c} &= \frac{\mathscr{X}_c}{W_c} \\ \dot{\mathscr{W}}_c &= \frac{\mathscr{W}_c}{V_c} \end{aligned} \tag{6}$$

2.2 Results

Figure 1 shows the scatter plot of $ln(V_t)$ versus $ln(\frac{V_c}{v_c})$ for the tests rated *Intermediate* (1) and *Advanced* (2). The different symbols used for these two ratings show quite a reasonable separation considering

$\mathscr{F}_w$	*Word Frequency*
$\mathscr{R}_w$	*Rareness* ($\propto \frac{1}{\mathscr{F}_w}$)
$\mathscr{W}_t$	*Tester Word Production*
$\dot{\mathscr{W}}_t$	*Tester Word Production Speed*
$\mathscr{X}_t$	*Tester Complexity*
$\dot{\mathscr{X}}_t$	*Tester Complexity Speed*
$\bar{\mathscr{X}_t}$	*Tester Complexity per Word*
$\mathscr{W}_c$	*Candidate Word Production*
$\dot{\mathscr{W}}_c$	*Candidate Word Production Speed*
$\mathscr{X}_c$	*Candidate Complexity*
$\dot{\mathscr{X}}_c$	*Candidate Complexity Speed*
$\bar{\mathscr{X}_c}$	*Candidate Complexity per Word*

Table 3: Nomenclature related to Qualitative Features

that only *Verbosity* information has been used. Figures 2- 4 show results for other possible pairs of tests and finally present a spread of all tests in this two-dimensional space for all three ratings. The separation seems much more pronounced between *Intermediate* and *Advanced* candidates than it is between Advanced and Superior using *Verbosity*. Of course, a non-linear clustering technique such as the one used for the OPIc in the second part of this paper should be able to provide us with considerable results based on the way the clusters look in the figures.

Figure 5 shows the scatter plot of the 44 tests in the training set rated *Intermediate* and *Advanced* by plotting $ln(\frac{v_c}{v_t}\frac{\bar{\mathscr{X}_c}}{\bar{\mathscr{X}_t}})$ versus $ln(\frac{\dot{W}_c}{W_t})$. Here, we have utilized both *quantitative* and *qualitative* features and the results seem very promising. In fact 91% of the data points making up the training data are linearly separable as demonstrated by the linear discriminator in figure 5. Figure 2 shows similar results between the 44 tests with ratings *Advanced* and *Superior*. In this case, a linear discriminant is no longer capable of doing considerable separation. Instead, ellipses have been drawn to demonstrate the capability of a set of Gaussian clusters in doing significant discrimination between the two classes. Figure 7 shows results of linear separation with 84% discriminability on the training data, between level 1 (*Intermediate*) and level 3 (*Superior*) tests.

Results seem quite promising and represent a clear answer of “yes” to the question posed in the title of this report. Further, more elaborate, studies are called for given these results.

2.3 Error Sources

There are few possible error sources in this process. Some important ones are listed here. First, there is the possibility of segmentation error. Manual evaluation of the performance of the segmentation for 10 randomly picked tests has shown an error rate of 2.86%.

Another source of error is identification error. Once the segmentation is completed, the candidate and the tester are identified. Quantitative results for this error rate are not available, but seem to be negligible. Although, an error at this level could profoundly affect end results since the candidate and tester will be mistaken for one-another.

Another source of error comes from the speech recognizer. The transcription error rate has not been quantified either. However, transcription accuracy for the purposes listed above is not as important as it would be in normal speech recognition applications.

Proficiency Level	Rating
Novice Mid	1
Novice High	2
Intermediate Low	3
Intermediate Mid	4
Intermediate High	5

Table 4: Ratings used in the OPIc tests

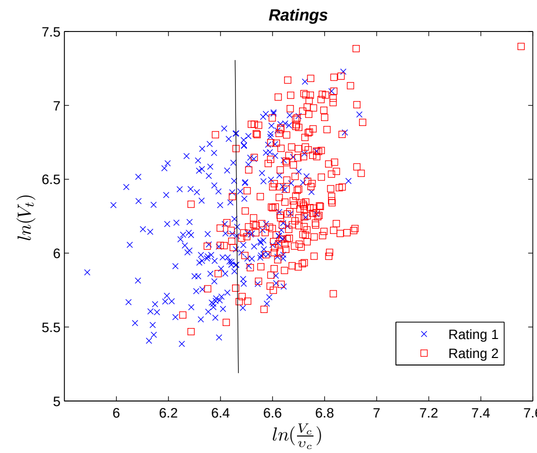

Figure 1: OPI Verbosity: Intermediate and Advanced

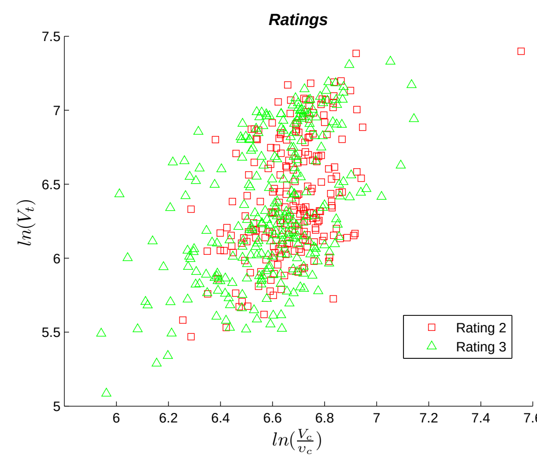

Figure 2: OPI Verbosity: Advanced and Superior

3 OPIc

In an OPIc scenario, there is no tester, or one may say that the computer acts like the tester. The candidate will make a self-assessment of his/her language proficiency. Then, a test is created for the candidate by combining a collection of Novice, Intermediate and Advanced prompts which are played back for the candidate and the candidate is expected to respond to them. These responses are recorded and used to rate the candidate's proficiency. Depending on the level of the prompt, there are limits to the number of seconds the candidate will be allowed to speak. These limits are made known to the following test process so that they may be used for normalizing the features being used in this part.

The audio data was recorded using the μ-Law amplitude coding technique [2] 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.[1] The compressed audio was uploaded to a server. The audio, in turn, was 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.

3.1 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. In this case, the *Verbosity* is computed by using the

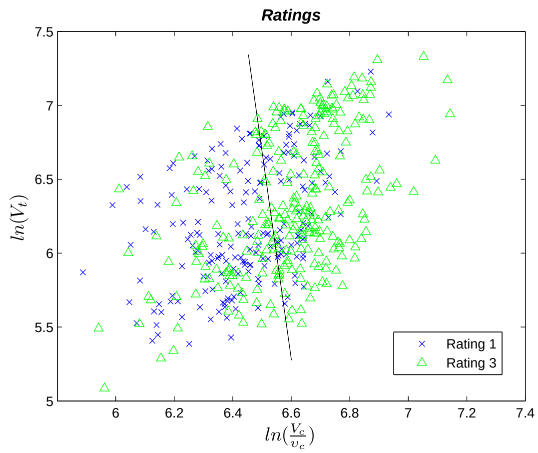

Figure 3: OPI Verbosity: Intermediate and Superior

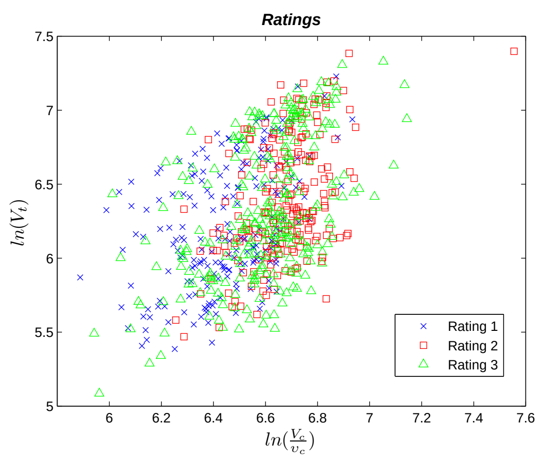

Figure 4: OPI Verbosity: Intermediate, Advanced and Superior

RecoMadeEasy$^{(TM)}$ engine of Recognition Technologies, Inc. to extract segments where audio is present. The length in number of seconds of spoken audio constitutes *Verbosity*. In the OPIc study, due to the lack of resources, only the *Verbosity* feature was used and *Rareness* and consequently *Complexity* features were not computed.

3.2 The Rating Process

In the OPIc case, an actual rating process was created and tested using the *Verbosity* feature. Let us assume that the feature for the i^{th} response is denoted as $\mathbf{f}_i$. Also, let l_k denote the k^{th} label. Possible labels in the OPIc study have been presented in Table 4. Theoretically, it is possible to describe any complex distribution by an infinite number of Gaussian distributions. However, in practice, this number may be finite to achieve a good approximation. If we assume that there exist a certain number of Gaussian Prototypes, the mixture of which describes the distribution of the features, then the a-posteriori probability of the label (rating) given a feature $\mathbf{f}_i$ may be estimated by equation 7.

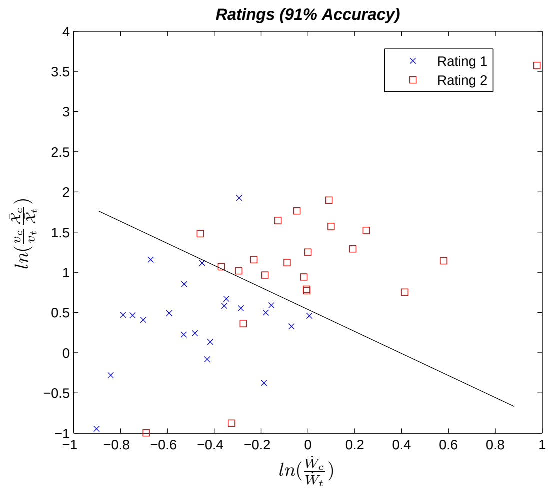

Figure 5: OPI Features: Intermediate and Advanced

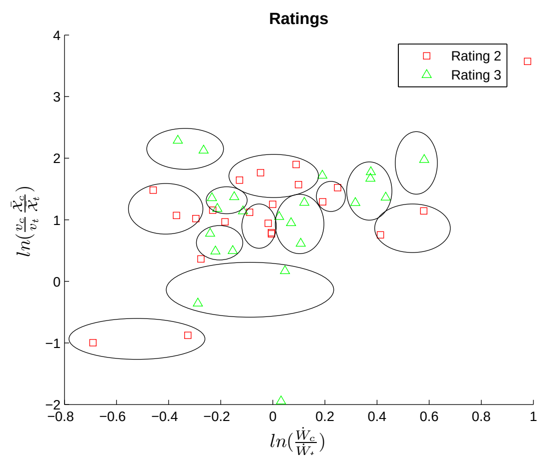

Figure 6: OPI Features: Advanced and Superior

$$p(l_k|\mathbf{f}_i) = \sum_{j=1}^{N} p(l_k|g_j)P(g_j|\mathbf{f}_i) \tag{7}$$

Then the *a-posteriori* probability of the Gaussian Cluster, g_i given the feature $\mathbf{f}_i$ is given by equa-

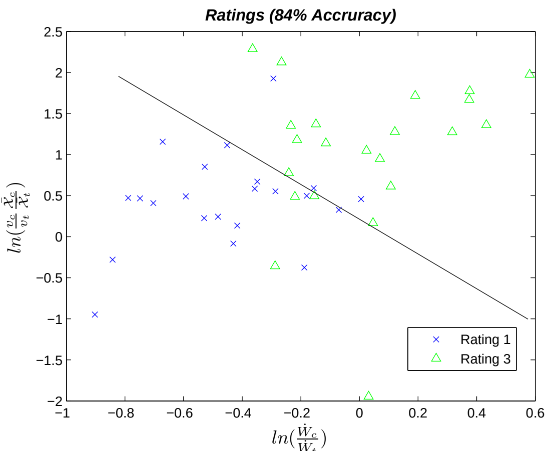

Figure 7: OPI Features: Intermediate and Superior

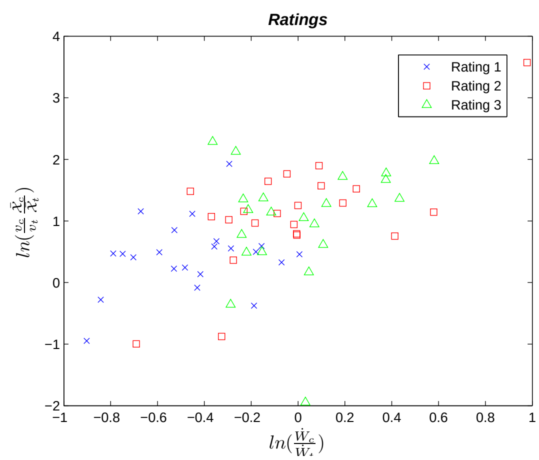

Figure 8: OPI Feature: Intermediate, Advanced and Superior

tion 8.[3]

$$P(g_j|\mathbf{f}_i) = \frac{p(\mathbf{f}_i|g_j)P(g_j)}{p(\mathbf{f}_i)} \tag{8}$$

$P(g_j)\ \forall\ j = 1,2,\ldots,N$ are the set of prior probabilities computed by the clustering technique. $p(\mathbf{f}_i)\ \forall\ i = 1,2,\ldots$ are assumed to be 1 since at any instance, i, this represents the probability of occurrence of feature $\mathbf{f}_i$, but since that feature is present at that moment, its probability is 1.

$p(\mathbf{f}_i|g_j)$ is also the likelihood of $\mathbf{f}_i$ given Gaussian g_j and may be computed using the equation for the Normal distribution,

$$p(\mathbf{f}_i|g_j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{f}_i - \boldsymbol{\mu})\right\} \tag{9}$$

$$\textit{where} \begin{cases} \mathbf{f}_i,\ \boldsymbol{\mu} \in \mathscr{R}^d \\ \mathbf{\Sigma} : \mathscr{R}^d \mapsto \mathscr{R}^d \end{cases}$$

In 9, $\boldsymbol{\mu}$ is the mean vector where,

$$\boldsymbol{\mu} \triangleq \mathscr{E}\{\mathbf{f}_i\} \triangleq \int_{-\infty}^{\infty} \mathbf{f}_i\ p(\mathbf{f}_i)d\mathbf{f}_i \tag{10}$$

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

$$\mathbf{\Sigma} \triangleq \mathscr{E}\left\{(\mathbf{f}_i - \mathscr{E}\{\mathbf{f}_i\})(\mathbf{f}_i - \mathscr{E}\{\mathbf{f}_i\})^T\right\} \tag{11}$$

$$= \mathscr{E}\left\{\mathbf{f}_i\mathbf{f}_i{}^T\right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{12}$$

Consequently, we may compute the likelihood of for any rating given the selected feature. 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.

3.3 Results

All together, there were 200 tests in each category but *Novice Mid* which only had 100. Three quarters of the data was used for training and the other quarter was used for testing. Figures 9 and 10 show the results of this algorithm applied to the training and test data respectively. Results have been obtained by changing the number of Gaussian clusters from 5 to 200. Results show that using this very crude feature, *Verbosity*, a respectable percentage in the mid-40s can be achieved for the middle-ground candidates. However, there is always a 0% accuracy for *Novice Mid* candidates. Also, the results are not as good for *Intermediate High* candidates.

Figure 11 shows a graph of the standard deviation from the ground truth which is a measure similar to the standard deviation from a mean where the mean is replaced by the ground truth of the label. This shows that in general grave mistakes are not made and most of the time the error is within one rating point away from the truth.

3.4 Error Sources

Microphone quality, background noise levels and microphone grounding issues have shown to be quite important in maintaining a good estimate for the quantitative features used here. Also, once qualitative features start to be used in the future, these factors may even play a bigger role. It is important to control these factors by trying to use a consistent microphone which is preferably connected through a USB connection with a built-in sound module so that grounding and other analog effects do not exist. Also, the **HE-AAC** format has a very aggressive compression which may affect the effectiveness of the qualitative features. These effects should be studied isolated in future studies.

Figure 9: OPIc Training Data Performance

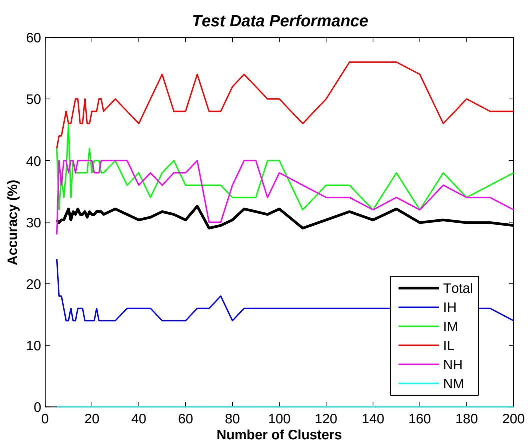

Figure 10: OPIc Test Data Performance

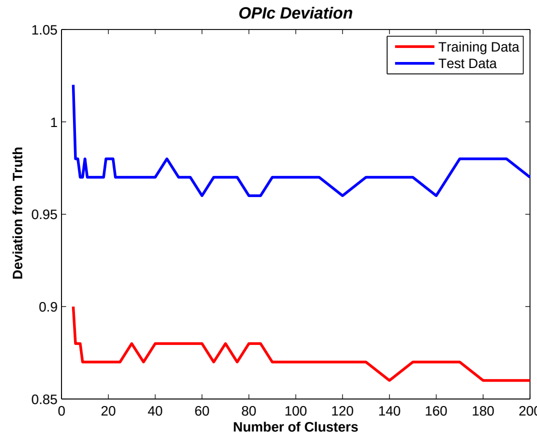

Figure 11: OPIc Deviation from Truth

4 Future Direction

- Improve the segmentation accuracy.
- Improve the identification accuracy.
- Try using combined Segmentation and identification.
- Increase the number of training and test samples for OPI by running all 726 files through the speech recognizer.
- Apply a degenerate single-state Hidden Markov Model approach to the OPI problem and study the accuracy across different number of clusters.
- Run the OPIc data through the speech recognizer to get the quality-type feature for the OPIc tests. If OPI is any indication, we should see significant improvements to the accuracy by adding this feature.
- Studies related to effects of noise.

5 Conclusion

It has been shown that using combinations of quantitative and qualitative features as defined in this research provide very promising capabilities for the automatic rating of candidates taking the OPI exams. Further, we have shown that for certain levels to be discriminated, a non-linear (Gaussian Mixture

Model) mimicking a single-state degenerate Hidden-Markov Model should be able to provide a much better performance. Also, through studying the OPIc exams, we have shown that automatic rating intermediate level exams is possible even with a very crude quantitative feature.

Combining the results obtained from the OPI and OPIc projects, it is quite apparent that in both cases, the quantitative and qualitative features as discussed here should be used in conjunction with a non-linear discriminative technique, namely the Gaussian Mixture Model (GMM) presented in this report. At this point each study was lacking a crucial part of the complete solution. In case of the OPI exams, only features were identified, but they were not used to do the final rating. In the OPIc case, although the discriminative technique showed promise, no qualitative feature was used and if OPI is an indication, there may be a large improvement expected by adding qualitative features to in the future.

Noise and microphone effects as indicated in the section entitled, "Error Sources", should be considered in future studies. The objective is to separate these and other effects by doing more tests so that contributions from different error sources are better quantified.

As a whole, the answer to the question of "whether Computer Analyses Can Predict Human Ratings of Speaking Proficiency" seems to be definitely affirmative.

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. The author would also like to thank the Human Language Technologies department of IBM Research for making its speech recognizer available to Recognition Technologies.

References

[1] Homayoon Beigi, *Fundamentals of Speaker Recognition*, Springer, New York, 2011, ISBN: 978-0-387-77591-3.

[2] G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," ITU-T Recommendation, Nov. 1988.

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

Homayoon Beigi, holds three employment positions. For the past few years, he has conducted research and development in the fields of Biometrics, Pattern Recognition and Internet-Commerce. As the President of Recognition Technologies, Inc., he is conducting research toward the production of a series of Speaker Recognition, Language Modeling and Signature Recognition line of RecoMadeEasy$^{(TM)}$ software engines. His work as the Vice President of Internet Server Connections, Inc. includes the development of the multiple-award winning Commerce Made Easy$^{(R)}$ software, an elaborate Electronic Commerce system used world-wide. In addition as an Adjunct Professor of Mechanical Engineering, he has taught "Applied Signal Recognition and Classification", "Speech and Handwriting Recognition" and "Digital Control" at the mechanical engineering and electrical engineering departments of Columbia University. He was a Research Staff Member at the IBM T.J. Watson Research Center from 1991 to 2001 where he conducted research on Speaker Recognition, Language Modeling, Aggressive Search Techniques, Speech Recognition, On-Line Handwriting Recognition, Software Architecture, Control Theory and Neural Network Learning.