Examining the Consistency of Evaluations Provided by Three Automatic Speech Recognition Systems

Howard Hao-Jan Chen
Department of English
National Taiwan Normal University

Abstract

In an EFL setting like Taiwan, it is often difficult to provide individual oral language training due to limited human resources. Recently, several innovative commercial CALL programs (e.g., TeLLmeMore and MyET) claim that they can provide high-quality pronunciation training with their automatic speech recognition technologies. Although these programs can provide scores and feedback to individual learners, many teachers and students are not sure if the evaluations provided by these systems are fair and consistent. In this study, the consistency of evaluations provided by three different ASR programs (TeLLmeMore, MyET, and Microsoft) was examined. Four different groups of college students were asked to interact with these different ASR systems. The subjects read the same 20 sentences to each system and then received scores and feedback from these three systems. The whole interaction processes and subjects’ performance and scores were recorded by Camtasia, a screen and audio capture tool. The audio files of the subjects’ performance were also evaluated by two human raters. Based on the scores assigned by human raters and ASR programs, strong correlations were found between human raters and different ASR scores. Moreover, strong correlations were also found among different ASR systems. These statistical results show that automatic speech recognition technologies can indeed assign consistent scores to students at different proficiency levels. Although at this stage these ASR systems might not be suitable for high-stake tests like TOEFL or college entrance examinations, these tools can be used for pedagogical interventions and low-stake tests like placement tests or diagnostic tests.

Key Words: Computer assisted language learning, pronunciation, automatic speech recognition, scoring, and feedback.

INTRODUCTION

Traditionally, English education in Taiwan puts much emphasis on developing students’ reading and writing skills. Listening and speaking skills thus did not receive much attention. Many college students indicate that they need to strengthen their listening and speaking skills. In the EFL (English as a Foreign Language) setting, speaking skill is commonly considered to be one of the most difficult skills for learners to improve.

In Taiwan, it is often difficult to provide high-quality oral language training since the average English class size is about 40-50 and most teachers are non-native English speakers. Moreover, most language programs at the college level still focus on teaching reading skills. Many college students find it very difficult to improve their oral communication abilities under these conditions.

However, the needs for better English oral ability are strong. Many universities and institutes worldwide ask students to demonstrate their oral proficiency before they are allowed to enter the English-speaking universities. Moreover, English language proficiency tests like GEPT (General English Proficiency Test) and TOEFL also help to push Taiwanese college students to further improve their speaking abilities.

The GEPT has two sub-tests: The first test focuses on reading/listening, and the second test
focuses on speaking/writing. The GEPT test-takers have to pass the speaking/writing tests before
they can receive a certificate. In Taiwan, more and more high school and college students are
required to take GEPT (General English Proficiency Test). Nevertheless, many students pass the
first test but fail in the second test.

The TOEFL test also begins to test student’s oral ability. In the next generation TOEFL test,
students will be asked to demonstrate how well they can use English to communicate by reading,
writing, listening and speaking. The new TAST (TOEFL Academic Speaking Test) is one
component of the completely revamped TOEFL test becoming operational in 2005. Until then,
students may practice taking the TAST privately using a computer and telephone. According to ETS,
one of the biggest challenges for North American academic institutions is determining if prospective
non-native English speaking students have an acceptable level of oral English language ability.
ETS expects that TAST can measure the ability to speak English clearly and fluently using situations
students may encounter in a real academic setting.

Based on the trends of language teaching and language testing in Taiwan and in the United
States, it is clear that English oral abilities should gain more and more attention.

**Improve Pronunciation with Automatic Speech Recognition Technologies**

Given that students have a strong need to improve their oral abilities, universities and institutes
have begun to explore different possible solutions. For instance, some universities hire more native
English speakers and ask students to interact with these teachers. Some colleges have reduced the
class size and expect that there will be more teacher-student interactions in the target language.
Some universities are also exploring the power of new computer technologies. There are indeed
several possible ways we can help college students improve their English pronunciation and
intonation.

Eskenazi (1999a; 1999b) proposed a successful computer-assisted pronunciation training
program should be able to provide the following conditions for language learners.
1. Learners must produce large quantities of sentences on their own.
2. Learners must receive pertinent corrective feedback.
3. Learners must hear many different native models.
4. Prosody (amplitude, duration, and pitch) must be emphasized.
5. Learners should feel at ease in the language learning situation.

For the five conditions stated above, most CALL (computer assisted language learning)
pronunciation training programs can accomplish point 1, 3, 4, and 5. But the second point is not so
easy to accomplish. Many pronunciation training programs can not process students' output and
provide adequate feedback. However, the scoring and feedback mechanisms are the keys to a
successful computer-assisted pronunciation training program.

More recently, automatic speech recognition technologies were used to provide scores and
feedback to language learners as the speech recognition technologies become more and more mature.
The technologies are widely used in various language teaching and learning programs. (Harless et al.
CALL programs based on different speech recognition technologies clearly have rather different mechanisms and options for offering feedback. Given the importance of scoring and feedback in any computer assisted language learning programs, it is necessary to examine these different scoring and feedback mechanisms more carefully.

In the sections below, we first introduce three CALL (Computer assisted language learning) products based on automatic speech recognition (ASR) technologies and briefly discuss their potentials for developing better pronunciation and intonation. These popular ASR programs are as follows: TeLL me More Pro, My English Tutor (MyET), and Microsoft Speech system. A detailed review of some other popular speech recognition programs is available in Chen (2001).

TellMeMore is a product from a French company, Auralog. Auralog develops and distributes software solutions and related services for foreign language learning. In TellMeMore Pro, each lesson follows the same format. Lessons are based on video scenes and include an interactive dialogue, a pronunciation practice, a comprehension activity and a range of additional exercises which support vocabulary and grammar structure acquisition. In the dialogue mode of this program, learners first listen to the program then choose one of three acceptable responses given on screen and pronounce it clearly into the microphone, as shown below in Figure 1. TellMeMore pro acknowledges the response by highlighting it. The dialogue develops according to the responses learners choose. If a response is not understood, learners must try again.

![Figure 1 TeLLmeMore Pro](image)

For the pronunciation practice, there is a 7-point scale scoring system. Learners will receive a score when they imitate a phase or sentence recorded by a native speaker. The scores are shown in Figure 2.
MyET (My English Tutor) is a new product developed by a team of Taiwanese speech recognition researchers. A typical screenshot is shown below in Figure 3. MyET is designed to be a personal tutor that can help learners learn English through speaking practice. MyET’s unique technology, "Automatic Speech Analysis System" (ASAS©), can analyze learners’ English speech on pronunciation, pitch, timing and emphasis, and even pinpoint problems to individual sounds. After listening to native speakers’ models, learners can try to imitate the native speakers’ pronunciations and intonations. After pronouncing the target sentence clearly into the microphone, they receive scores and immediate feedback on how to improve.

Another very important feature of MyET is that it is completely web-based. All the interactions with the ASR program are carried out via the Internet connections. This feature also allows the company to host online learning communities, and students can compete with each other in various virtual learning communities.

Microsoft has been developing Microsoft Speech .Net technologies for several years. The SASDK (Speech Application SDK) and Speech Server (Speech Engine Services) provided by Microsoft can be used to develop both voice-only (Telephony) and multimodal (desktop pc, notebook pc, pocket pc, other mobile devices) speech recognition/synthesis environment. We can use SASDK and Visual Studio ASP.NET to edit web pages (HTML + SALT + JScript) and put these on the Web Server (IIS), and the Client side can easily use web browser to read the web page which
contains speech tags. Web Server will process the voice input from the client or send the TTS (Text to Speech) results to the client. On desktop or notebook computers, users can download and install the Speech add-in for Microsoft Internet Explorer to personal computers. After installing the software, the computer will have the speech API for speech recognition and the TTS prompt engine for speech synthesis. The users only need to use Internet Explorer to read the web pages (HTML + SALT + JScript), and speech add-in will automatically decode SALT (Speech Application Language Tags) grammar and perform speech recognition and speech synthesis on the desktop or notebook computers.

Based on the Microsoft speech technologies, we developed a web-based speech recognition system for practicing English pronunciation. This ASR web site allowed students to practice their pronunciation and oral skills. Students can gain access to this web from anywhere at any time. When students log onto the web site, they will follow the instructions to install the Microsoft speech add-in for Internet Explorer. After installing the speech add-in, students can begin to interact with the ASR web site. ([http://140.122.83.197/test](http://140.122.83.197/test) or [http://140.122.83.197/pronunciation](http://140.122.83.197/pronunciation)) A preliminary survey indicated that college students enjoy using several practice units available from the web site.

One exercise called “Listen and Repeat” is shown below in Figure 4. The type of exercises asks students to listen and repeat the sentence shown on the screen. This exercise is a very popular test format used in several spoken English tests (e.g., Ordinate Phonepass or GEPT tests). We can put in many commonly used sentences into the system. Students can listen to the native speakers’ pronunciations first and then begin to practice their pronunciations. After they repeat the sentence, the system gives immediate feedback on their performance. If the answer is acceptable, a loud applause follows and system marks that the learner has passed the item. If the answer is not acceptable, then the system shows “no recognition”. Students can then try again.

![Figure 4 Listen and Repeat in The Microsoft ASR system](image)

**Comparing ASR Scoring and Feedback Mechanisms**

Generally speaking, scoring can be done on the basis of comparison between temporal properties (e.g., rate of speech) and/or acoustic properties of the student’s utterance on one side, and native’s reference properties on the other side: the closer the student’s utterance comes to the native models used as reference, the higher the score will be.

These different types of scores and feedback provided by the systems could able to help
learners understand the current level of their speaking skills and let them know when they are improving. (cf. Cucchiariini et al., 2000a, 2000b, 2000c; Neri et al., 2001, 2002, 2003) The scoring and feedback mechanism of the three different ASR programs can be summarized in the following table.

Table 1
The comparison of the Scoring and Feedback Mechanism in Three ASR Programs

<table>
<thead>
<tr>
<th>Scoring mechanisms and Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>TeLL me More Pro</td>
</tr>
<tr>
<td>For pronunciation exercises, 1-7 points are awarded based on learners’ performance. For video-based exercises, no specific scores were assigned to students. The program either accepts learner’s output or asks learners to try again.</td>
</tr>
<tr>
<td>MyET</td>
</tr>
<tr>
<td>A total score and four sub-scores (pronunciation, pitch, timing and emphasis) are generated for each utterance. The scores are presented to learners and more detailed feedback for each sub-section can be provided upon request. (full score:100)</td>
</tr>
<tr>
<td>Microsoft Speech Engine</td>
</tr>
<tr>
<td>Different exercises can be created with this system. The Listen and Repeat Exercise either allows learners to move on (if learner’s output is acceptable) or asks learners to try again. (no specific score is assigned)</td>
</tr>
</tbody>
</table>

TellMeMore provides two types of scores and feedback: For video-based conversations, if students produce acceptable output, they are allowed to move on to the next exercise. If they fail to reach the standard, they are asked to try again. For the pronunciation exercises, they receive a score (1-7 point scale) right after they read a sentence or phrase assigned by the program. MyET program provides scores for each sentence read by learners. Students receive one total score and four sub-scores for their performance in pronunciation, pitch, timing and emphasis. MyET program can further analyze students’ English speech on pronunciation, pitch, timing and emphasis, and pinpoint problems of individual sounds. Microsoft Speech system only checks if learners’ performance has reached the threshold level and asks learners to go on or to try again. Its feedback is similar to the video-based exercise of the TellMeMore program; there is no specific score. We can only count how many questions were marked “successful” or “unsuccessful”.

Research Questions

As shown above, there are several different ASR programs available for language teachers and learners; however, we are not sure if they can provide consistent assessment. Can these ASR programs successfully differentiate between orally more proficient students and less proficient students? Are the scores assigned to students by various ASR programs comparable to the scores assigned by human raters or language teachers? In addition, teachers and students also wonder if different speech recognition systems can generate consistent evaluations (similar scores). Which automatic recognition engine will provide better assessment for non-native speakers of English? Can they provide similar scores to a certain group of students?

Given the importance of accurate scoring and feedback in any ASR-based pronunciation
training programs, it is essential to conduct an empirical study to closely examine if ASR programs can provide fair and consistent assessment comparable to human raters. These questions can be answered if we compare the scores assigned by human raters and the scores offered by different automatic speech recognition programs. We can examine the correlations between the rater scores and computer ASR scores and also the correlations among different ASR scores. These comparisons will help us determine if these scores and feedback from different ASR programs are consistent.

**METHODS**

**Participants**

We need subjects from different proficiency levels to determine if ASR programs can assign appropriate scores to students at different proficiency levels. 53 college students at National Taiwan Normal University (NTNU) participated in this study. These students came from various departments at NTNU. Group A subjects were freshmen from the English Department, Group B subjects were seniors from the English department. Group C subjects were freshmen taking general English course, and they were from the high intermediate classes. Group D subjects were also freshman taking required English course, but they were from the lower level classes.

Group C and Group D subjects were from different departments of NTNU and their English proficiency levels were based on their English scores of the JCEE (Joint College Entrance Examination). Group A and B both have higher English proficiency because they were selected and trained to be English teachers. The subjects in these two groups all gain very high English scores in the joint college entrance examination. Group A has 10 subjects, Group B has 20 subjects, and Group C has 11 subjects, and Group D has 12 subjects.

**Materials and Procedures**

The four groups of subjects were invited to interact with the three ASR programs in a computer lab. Two research assistants were present to help individual learners interact with these three programs.

To assess the consistency of these three programs, we loaded the same set of sentences to each of these three ASR programs. Since we cannot freely load sentences into the TellMeMore system (a CD-ROM based program), we decided to use the existing pronunciation exercise in TellMeMore and load the same set of sentences onto the other two systems.

One unit of TellMeMore pronunciation exercise, which contains twenty sentences, was used as the test (Appendix I). The subjects had 3 minutes to preview these 20 sentences and read each of these twenty sentences to three different ASR programs via microphones attached to computers. Subjects interacted with these three ASR systems and received scores and feedback from these systems. The order of using these three programs was randomly assigned. All the process of interaction was recorded by a screen and audio capturing program called Camtasia. The program was able to record student’s voices and also scores and feedback shown on the computer screen. All these digital files were saved for later transcription and analysis.
Rating

Subjects’ performances were evaluated by both ASR programs and the human raters. TellmeMore assigned 1-7 points to subjects based on their performance. MyET assigned 1-100 points to subjects based on their performance. Microsoft did not assign scores but either accepts or rejects subjects’ performance. The audio files of these subjects’ performance were also saved and evaluated by two human raters. Two human raters are full-time graduate students in TESOL and Linguistics programs. The two raters discussed the rating scales with the project researcher before grading subject’s performances. A 10-scale rating system was adopted to assign scores to these subjects.

In the following sections, we first present the scores assigned by human raters and the scores assigned by three speech recognition programs. Then we examine the correlations between the scores assigned by human raters and the scores generated by three different ASR programs. Last, the correlations among three different ASR programs were also examined.

RESULTS

Human Rater Scores

The scores assigned by two different human raters were summarized below in Table 2. The mean score of Group A is 7.1. The mean score of Group B is 7.5. The mean score of Group C is 4.8. The mean score of Group D is 3.2. Based on the scores, it is clear that subjects from the English department gained higher scores. The high intermediate level subjects also performed better than lower level subjects. Learner performances are closely related to their English proficiency levels.

<table>
<thead>
<tr>
<th></th>
<th>Group A (n=10)</th>
<th>Group B (n=20)</th>
<th>Group C (n=11)</th>
<th>Group D (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater A</td>
<td>8.0</td>
<td>8.0</td>
<td>5.90</td>
<td>4.17</td>
</tr>
<tr>
<td>Rater B</td>
<td>6.2</td>
<td>7.0</td>
<td>3.72</td>
<td>2.16</td>
</tr>
<tr>
<td>Average</td>
<td>7.1</td>
<td>7.5</td>
<td>4.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>

To test if the two raters agreed with each other in assigning scores to subjects, a Pearson correlation coefficient test was conducted by using SPSS program to determine if any correlation existed between the scores from two human raters. It turned out that there was a strong positive correlation between two raters ($r = .870**$, $p < .01$).

The one-way ANOVA analysis indicated that there was a significant difference among these four groups, as shown below in Table 3. A post-hoc Scheffe test, shown below in Table 4, further indicated that there was no significant difference between Group A and Group B. However, there were significant differences among Group A, Group C and Group D. There were also significant differences among Group B, Group C and Group D. C Group was significantly different from Groups A, B, and D. Group D was also significantly different from Groups A, B, C.
Table 3
One-Way ANOVA Comparing Means of Group Scores by Human Raters

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3</td>
<td>169.655</td>
<td>56.552</td>
<td>53.120*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>49</td>
<td>52.166</td>
<td>1.065</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>211.821</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P < .000

Table 4
Between Group Comparison of Means of Scores by Human Raters

<table>
<thead>
<tr>
<th>Mean (%)</th>
<th>Group</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Group A</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>Group B</td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>4.8</td>
<td>Group C</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>3.2</td>
<td>Group D</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level

Computer ASR Scores

The average scores assigned by three different ASR program were summarized in Table 5.

Table 5
Total Scores of These Three Different Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>TellMeMore Pro (full score: 7)</th>
<th>My English Tutor (full score:100)</th>
<th>Microsoft Speech Engine (full score:20)(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>5.40</td>
<td>83.10</td>
<td>15</td>
</tr>
<tr>
<td>Group B</td>
<td>6.05</td>
<td>87.80</td>
<td>15.4</td>
</tr>
<tr>
<td>Group C</td>
<td>5.17</td>
<td>81.64</td>
<td>12.4</td>
</tr>
<tr>
<td>Group D</td>
<td>3.65</td>
<td>73.67</td>
<td>6.66</td>
</tr>
</tbody>
</table>

TellMeMore Scores

The scores assigned by TellMeMore are given in Table 5. The mean score of Group A is 5.40. The mean score of Group B is 6.05. The mean score of Group C is 5.17. The mean score of group D is 3.65. Similar to what was found from human raters, English majors consistently gained better scores. The general patterns (B > A > C > D) generated by TellMeMore is consistent with the rank order provided by two human raters. The one-way ANOVA analysis further indicated that there

\(^1\) The full score is 20 because there are only twenty sentences. If the system accepts what subjects say, they will gain one point for one sentence.
was a significant difference among these four groups, as shown below in Table 6. A post-hoc Scheffe test, shown below in Table 7, further indicated that there was no significant difference between Group A, Group B, and Group C. However, Group D was significantly different from Group A, Group B and Group C.

Table 6
One-Way ANOVA Comparing Means of Group Scores by TellMeMore Program

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3</td>
<td>43.959</td>
<td>14.653</td>
<td>15.053*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>49</td>
<td>47.698</td>
<td>.973</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>91.656</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P < .000

Table 7
Between Group Comparison of Means of Scores by TellmeMore Pro Program

<table>
<thead>
<tr>
<th>Mean (%)</th>
<th>Group</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.40</td>
<td>Group A</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>6.05</td>
<td>Group B</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>5.17</td>
<td>Group C</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>3.65</td>
<td>Group D</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level

MyET Scores

The scores assigned by MyET are given in Table 5 shown above. The mean score of Group A is 83.10. The mean score of Group B is 87.80. The mean score of Group C is 81.64. The mean score of group D is 73.67. Similar to what was found from human raters and the TellMeMore program; English majors consistently gained better scores. The general patterns (B > A > C > D) generated by MyET is consistent with the rank order provided by two human raters and by TellMeMore.

The one-way ANOVA analysis further indicated that there was a significant difference among these four groups, as shown below in Table 8. A post-hoc Scheffe test, shown below in Table 9, further indicated that there was no significant difference between Group A and Group B. Group A significantly outperformed Group D only. Group B significantly outperformed both Group C and Group D. Group C was significantly different from Group B and Group D.
Table 8
One-Way ANOVA Comparing Means of Group Scores by MyET

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3</td>
<td>1505.133</td>
<td>501.771</td>
<td>27.318*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>49</td>
<td>899.925</td>
<td>18.366</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>2405.058</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P < .000

Table 9
Between Group Comparison of Means of Scores by MyET

<table>
<thead>
<tr>
<th>Mean (%)</th>
<th>Group</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>83.10</td>
<td>Group A</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>87.80</td>
<td>Group B</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>81.64</td>
<td>Group C</td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>73.67</td>
<td>Group D</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level

Microsoft Speech System Scores

The scores assigned by Microsoft Speech System is given in Table 5 shown above. The mean score of Group A is 15. The mean score of Group B is 15.4. The mean score of Group C is 12.4. The mean score of group D is 6.66. Similar to what was found from human raters and the TellMeMore and MyET programs; English majors consistently gained better scores. The general patterns (B > A > C > D) generated by the Microsoft system is again consistent with the rank orders provided by two human raters, TellmeMore, and MyET.

The one-way ANOVA analysis further indicated that there was a significant difference among these four groups, as shown below in Table 10. A post-hoc Scheffe test, shown below in Table 11, further indicated that there was no significant difference between Group A, Group B, and Group C. These three groups, however, significantly outperformed Group D.
Table 10  
One-Way ANOVA Comparing Means of Group Scores by Microsoft Program

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>D.F.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>3</td>
<td>635.271</td>
<td>211.757</td>
<td>15.395*</td>
</tr>
<tr>
<td>Within Groups</td>
<td>49</td>
<td>674.012</td>
<td>13.755</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>1309.283</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P < .000

Table 11  
Between Group Comparison of Means of Scores by Microsoft Program

<table>
<thead>
<tr>
<th>Mean (%)</th>
<th>Group</th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Group A</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.4</td>
<td>Group B</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.4</td>
<td>Group C</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.66</td>
<td>Group D</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

* The mean difference is significant at the .05 level

The Correlations between Human Rater Scores and Computer ASR Scores

Based on the statistical analyses, we found that the scores from both human raters and three different ASR programs were very similar and the rank orders by these four types of raters (both human rater and ASR recognizers) are very consistent. In addition to the overall ranking order, we are also curious if there is any correlation between human raters’ scores and the three computer ASR scores. How strong was the correlation? Based on the four different types of scores, we found there were strong correlations between human rater scores and ASR scores. As shown below in Table 12, the correlation between human rater scores and the TellmeMore score is high (r=.708 **, P< .01).

Table 12  
Correlation Matrix for Human Rater Scores and TellMemore Scores

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>TellmeMore</td>
<td>.703</td>
<td>.672</td>
<td>.708**</td>
</tr>
</tbody>
</table>

p < .01, N=53.

As for human rater scores and MyET scores, again we also found strong correlation (r=.784 **,
The correlation is even stronger than that of the TellmeMore program.

Table 13
Correlation Matrix for Human Rater Scores and MyET Scores

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyET</td>
<td>.768</td>
<td>.752</td>
<td>.784**</td>
</tr>
</tbody>
</table>

$p < .01, N=53.$

As for human rater scores and Microsoft system scores, we also found strong correlation ($r=.708 **, P< .01$). The correlation between human raters and computer ASR scores is also strong.

Table 14
Correlation Matrix for Human Rater Scores and Microsoft Scores

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>.669</td>
<td>.697</td>
<td>.708**</td>
</tr>
</tbody>
</table>

$p < .01, N=53.$

Based on the statistical analyses, it is clear that the human rater scores have strong correlations with different computer ASR scores. It seems that the ASR program can provide similar scores comparable to the scores assigned by human raters.

The Correlations among the Computer ASR Scores

In addition to the correlation between human rater scores and ASR scores, we also examined the possible correlations among scores provided by the three different speech recognition systems. How strong is the correlation among these different ASR programs?

Again, there are strong correlations among the different ASR scores (TellmeMore and MyET, .789; MyET and Microsoft, .555; TellMeMore and Microsoft, .726). There are strong correlations among these different scores. These results show that different automatic speech recognition technologies can indeed assign rather consistent scores to subjects at different proficiency levels.

DISCUSSIONS

The results in previous sections show that there is a significant positive correlation between the
human raters and Speech recognition programs (r= .708 respectively, P <0.01). It is clear that there is a strong correlation between human rater scores and the computer ASR scores. Moreover, it is also important to note that the group ranking by either human raters or different ASR program is rather consistent (Group B > Group A > Group C > Group D).

In addition to the strong correlations between human and ASR programs, the correlations among the three different ASR programs are also strong. For TellmeMore Pro, we found that it has a strong positive correlation with both MyET and the Microsoft Speech Engine. Though the correlation between MyET system and Microsoft Speech has not reached the level of strong correlations (r= .555, P <0.01), the correlation is still statistically significant.

These positive results from different types of comparison show that the state-of-the-art automatic speech recognition can indeed be useful for language teachers and learners. Many schools and colleges in EFL settings want to help students improve their oral abilities, but they do not have enough qualified teachers and class time to help their students. The ASR technologies might be able to play an important role in these settings. In this study, the lower level students consistently gained lower scores from different ASR programs. We can use ASR programs to quickly identify students who might need extra help on pronunciations and oral communication. We can first use any of these ASR programs to find out these students who really need help and then use these ASR programs to help these students improve their pronunciation and oral abilities. Given the high correlations between human raters and ASR programs, administrators and teachers should feel more comfortable to allow learners to interact with these quality ASR programs.

We also noticed that the TellMemore program and MyET programs were more tolerant with lower level students. Lower level students can still get passing grade from MyET program (the average score is around 73.67). The main reason for this phenomenon might be related to different standard used by these three ASR systems. TellmeMore and MyET programs were mainly designed to deal with non-native speakers. As far as we know, TellmeMore and MyET in fact both used ESL/EFL learner speech as part of their training data.

The Microsoft speech system is different from the other ASR programs because it is mainly designed to process the spoken input from English native speakers. In this study, we also found that the Microsoft system is indeed a very sensitive system and its recognizer is stricter than the recognizers used by TellMeMore and MyET. Since Microsoft system only indicated it accepted or rejected learners’ performances, the lower level learners often felt rather frustrated when they only received rejections from the system. MyET system, on the other hand, will assign these learners with a better score- 70 or higher.

Although the MyET system can assign a more encouraging score to the lower level EFL learners, it also runs the risk of sending an incorrect message to these learners. Based on the scores, some of these learners might believe that their own pronunciation and intonation are not bad, and they do not need to worry about their oral abilities. Therefore, how to provide meaningful

2. As can be found in our study, English majors might not need these ASR programs since most of them can obtain fairly high scores from these ASR programs.
scores and proper feedback to different levels of language learners might be a very crucial research topic.

Even though the ASR programs might not be able to replace human raters for high-stake language tests, it seems possible that these new tools can be used for low-stake tests like placement tests and diagnostic tests.

**LIMITATIONS AND FUTURE RESEARCH**

Although this study confirms the usefulness of quality ASR programs and uncovers some difficult issues about scoring and feedback, there are still some limitations about this study. First of all, we have only included limited number of sentences (20 sentences and each sentence is only about 4-11 words), we should include more and longer sentences in the future research. Second, if possible, we should allow learners to try different production tasks. In this study, we only ask learners to read sentences aloud. Perhaps we can ask them to read words and phrases in the future research. Third, there were only 53 students participating in this preliminary study, more subjects at different proficiency levels should be included in the future studies.
Appendix I.
1 You would be acting as our representative.
2 That's true, you're quite right.
3 You were going to pay in the event of shared ownership.
4 Agreed, but the costs aren't very high, are they?
5 We're not going to block the deal because of this.
6 Any faulty article will be replaced.
7 Our company has a worldwide reputation;
8 we are renowned for the quality of our products.
9 We will allow a discount on any faulty item.
10 You decide that when you receive the goods,
11 provided the level is lower than 5%.
12 We can use the same quality criteria
13 that we use in our own factory.
14 The criteria are obvious.
15 What happens if we are unable to reach an agreement?
16 We should include a clause
17 providing for automatic renewal.
18 This is generally how we proceed.
19 If one of us wants to terminate the contract,
20 he notifies the other three months in advance.

ACKNOWLEDGEMENTS
This research was funded by the National Science Council (NSC 92-2411-H-003-039 and NSC 93-2411-H-003-012)

References


