NASA-Ames Voice Recognition and Response Evaluation Report

Forward by UFA, Inc.

UFA's ATVoice® Voice Recognition and Response (VRR) product was recently integrated with ATCoach®, the Embedded Test & Training Simulator for STARS, at Boston Consolidated TRACON. The FAA then commissioned an independent Evaluation Study to determine if VRR is acceptable for TRACON instructional purposes.

The VRR Evaluation Study was conducted by a team led by Richard Mogford, PhD of NASA-Ames Research Center. The VRR Evaluation Report summarizes the results from this study.

- This is the first, and the only known, formal study independently conducted evaluating the efficacy of Voice Recognition and Response technology in an Air Traffic Control training domain.

- The Evaluation Report makes the following significant points:
  - VRR is acceptable for TRACON instructional purposes.
  - The Report recommends the deployment of VRR to other STARS sites.
  - VRR performed well as compared to the existing Pseudo Pilot (PP)-based system.
  - VRR really came into its own in complex, high demand situations.
  - VRR may shorten the time required to develop exercises and may improve the quality and flexibility of FAA terminal training programs.
  - VRR errors tended to be less disruptive with easier recovery than PP errors.
  - VRR always acts in accordance with its read back; not so for PPs (e.g., command entry errors).
  - 90% and 110% capacity exercises were utilized in order to stress VRR.
  - Only two minor edits are required to existing exercises for VRR readiness.

As the evaluation focused on the efficacy of VRR, it did not evaluate other very important points that deserve mentioning regarding the application of VRR technology in this domain:

  - Major reduction in training costs are foreseen by eliminating or significantly reducing the number of PPs required.
  - Training is available 24 hours per day, 7 day per week.
  - VRR does not take leave or receive other assignments as humans do.
  - Approximately two-thirds of the integration effort addressed required improvements to the simulator to make it "voice-ready" / "response-ready". The integration of VRR is quite straight forward.
  - Controller and PP tend to train as a team helping each other through busy periods and recovering from errors. With VRR, Controllers are more careful with their phraseology and listen more intently to the read backs – a significant training advantage with VRR.
  - Using different Pilot voices for aircraft substantially increased the realism as compared to the same voice for all aircraft with PP-based systems.

With operations in MA and MD and two sites in Germany, UFA, Inc. is a global leader in simulation systems for the ATC and ATM markets. Its ATVoice product, acclaimed by many customers as the most advanced system available, is integrated with UFA's radar and tower simulators. ATVoice may also be integrated with third party simulators and other applications in and out of the air traffic sector. Contact Rajiv Sood at 301.216.2717 or soodr@ufainc.com.

Voice Recognition and Response Evaluation Report
Phases 1 and 2

Submitted by:
Richard H. Mogford, Ph.D.
NASA Ames Research Center
January 3, 2007

EXECUTIVE SUMMARY

The Federal Aviation Administration (FAA) is implementing new systems to enhance the performance of its training simulators. A recent project has been to add voice recognition and response (VRR) to the Standard Terminal Automation Replacement System (STARS) simulator. This may have the benefit of reducing training costs since human pseudopilots (PPs) might no longer be needed.

NASA Ames Research Center completed an assessment of the effectiveness of the STARS VRR system at the Boston Consolidated Terminal Radar Control (TRACON) facility. The VRR system was evaluated at two points in time and the results are reported as Phase 1 and Phase 2.

In Phase 1, VRR performed well as compared to the existing PP system. During the assessment runs, the errors tended to result in fewer disruptions to training than those in PP runs. However, some problems existed with voice recognition for specific users that needed to be rectified.

In Phase 2, there were fewer errors. The system did much better with voice recognition for users who had difficulty in Phase 1. Also, the type of errors changed from recognition of words that might be found in standard FAA phraseology to more specific, well-defined problems that could be rectified on the spot, or dealt with through future software changes.

The assessment exercise has been valuable for VRR system development. We recommend that the VRR developer continues with improvements to the VRR system to enhance its accuracy, based on the findings of this evaluation. We also recommend a Phase 3 test of the VRR system with developmentals. This is important since the VRR capability will frequently be used for developmental training and it has not yet been assessed with this group. VRR should also work adequately well with most experienced controllers for training purposes. However, efforts will be needed with some participants to help them use standard phraseology while using this system.

Assuming that improvements are made based on these findings, the system should be ready to deploy to other sites. Testing with developmentals could be completed at Boston Consolidated TRACON, or could be done at the next deployment location. It might be best to field the system in stages, and assess its success at a limited number of other facilities before proceeding further. When the VRR system is installed at other sites, there will no doubt be a need to make local adjustments and configuration changes. It will be important for on-site staff to work with the system and have the support of its developers to make adjustments, as needed.

TABLE OF CONTENTS

1.0 OVERVIEW

The Federal Aviation Administration (FAA) is implementing new systems to enhance the performance of its training simulators. A recent project has been to add voice recognition and response (VRR) to the Standard Terminal Automation Replacement System (STARS) simulator. This may have the benefit of reducing training costs since human pseudopilots (PPs) might no longer be needed. It also may shorten the time required to develop training scenarios and improve the quality and flexibility of FAA terminal training programs.

The FAA requested an assessment of the effectiveness of the STARS VRR simulation system. NASA Ames Research Center completed this evaluation at the Boston Consolidated Terminal Radar Control (TRACON) facility. The VRR system was evaluated at two points in time and the results are reported separately in this document as Phase 1 and Phase 2.

The evaluation was to determine if VRR for the STARS training simulator is acceptable for TRACON instructional purposes. These systems may be used for developmental and Certified Professional Controller (CPC) proficiency training. Candidate evaluation measures included development time, VRR errors, and impact of errors on training.

2.0 CONSTRAINTS

This study was not designed as a controlled experiment with sufficient data points to make statistical comparisons. It was also not possible to test developmentals at the test site, since none were available during the assessment period. CPCs participated in the data collection sessions. This made it difficult to extend the findings to developmentals.

3.0 PHASE 1

3.1 Method

3.1.1 Equipment

The research was conducted in the training simulator room at Boston Consolidated TRACON. There were eight STARS consoles in the training room, and two were configured with the VRR capability. A special audio jack box was installed on each console for headsets. There was air traffic control-specific logic in the VRR software that enhanced the voice recognition process. The system was speaker independent, so no voice training was needed. The baseline dictionary and grammar files supported standard phraseology per FAA Order 7110.65. There were four distinct pilot voices.

During VRR scenarios, one STARS console was used for the trainee. There was a 17-inch monitor in the console next to the trainee that showed information on the VRR system. The human PP who managed the simulated aircraft used a separate STARS display. When the PP made keyboard entries, they appeared in a preview area on their STARS display.

3.1.2 Participants

The participants were CPCs from Boston Consolidated TRACON who took time away from their normal duties to participate in the study. The assessment schedule (see Appendix A) included a total of 36 trials or runs with 12 controllers (24 runs in Phase 1 and 12 in Phase 2). We used two different PPs who normally assisted with training in the facility.

3.1.3 Scenarios

The scenarios for the evaluation were drawn from those actually used in training. Twelve scenarios were selected that varied in complexity and type. Moderate scenarios were at 90% of

normal traffic volume and complexity, while difficult scenarios were at 110% of normal volume and complexity. (Low volume/complexity scenarios were not used given that the goal was to stress the VRR system.) Scenario types included arrival, departure, and sector (en route) traffic. The scenarios also had different runway configurations. (See Appendix A for details of the schedule and scenario types.)

3.1.4 Experimental Design

The objective of the study was to run the PP and VRR systems through a range of conditions, varying participants, PPs, scenario types, and difficulty. The design for Phase 1 (as shown in Appendix A) had 24 trials or data points. Six controllers participated, with each controller running four trials each day. Controllers worked two traffic scenarios in the PP condition, and the same two scenarios in the VRR condition. For each controller, one scenario was moderate and the other was difficult.

3.1.5 Data Collection

We interviewed the Boston Consolidated TRACON instructor who worked with the VRR system to assess the nature of the differences in the tasks required to prepare PP versus VRR scenarios.

VRR data collection trials were run in the training simulator one participant at a time. This proved to be an efficient approach given the various kinds of observational data that were collected. Data collection forms were created that captured the number and type of errors (see Appendix B). The errors were categorized as shown in Table 1.

Table 1. PP and VRR error types.

Error Type	Error Code
No Response	N
Incorrect Readback	R
Pilot "Say Again"	P
Controller "Say Again"	C
Pilot Deviation (due to error executing clearance)	D
PP did not make entry	N
PP made wrong entry	I

There were four observers for each run, including an instructor from Boston Consolidated TRACON and three researchers. The instructor and two researchers kept track of PP or VRR errors. Due to the rapid pace of the scenarios, we found that using multiple observers helped collect valid data, since we could compare and corroborate our findings after each run. Another observer, using a separate form, noted whether the PP made data entry errors during PP runs.

The observers compared their error forms after each run and one form was corrected to reflect the consensus of observations from the group. The instructor made a "yes/no" assessment of each error to determine if it had serious negative consequences for the training session.

Voice recordings and data files were collected from the PP and VRR systems in case further analysis was needed.

In order to assess training effects, at the end of each trial, the participating controller was asked to fill out a questionnaire regarding the performance of the PP or VRR system. The instructor filled out a similar questionnaire. See Appendix B for the questionnaire forms.

3.2 Results

The procedures for preparing PP and VRR scenarios were reviewed with the instructor who had worked extensively with both systems. For PP or VRR, the simulator uses the same basic software (AT Coach) for running scenarios. Two types of minor changes are required to make a scenario "voice ready." First, frequencies for voice channels are entered into the facility's AT Coach Site File. This is a one-time edit and permits the use of multiple voice positions within any designed scenario. Secondly, the assignment of a voice channel to each desired aircraft is necessary within each Scenario File. This can be done by simply editing a text file. Scripting or special events may be added to AT Coach during a run, and this is accomplished in essentially the same way for both PP and VRR systems. Based on this, there appeared to be few differences in the scenario preparation time and workload between the PP and VRR-based simulation systems.

Our original plan was to conduct 24 trials of the PP system and compare this to 24 identical runs using VRR (for a total of 48 runs). As we progressed through the trials, it became evident that we had collected sufficient data by the end of the 24th run to draw some initial conclusions. We also noted that the VRR system had some problems recognizing some of the utterances of two of the controller participants, to the point where it was clear that improvements were needed before more testing was completed.

Figure 1 shows a graph of the total number of errors for PP and VRR runs. This includes all of the error types shown in Table 1. The PP and VRR made different kinds of errors, but all are plotted here. Plots are included for moderate and difficult runs, and for the different types of scenarios (arrival, departure, and sector).

The vertical lines on each bar of some of the bar charts are called “error bars” and show the amount of variability in the results. The length of the line is the amount that ratings vary between the respondents. Longer error bars are observed when there is a small number of people making ratings and their opinions differ. In several of the graphs, the error bars are large, indicating a disagreement in questionnaire ratings between the respondents. In such cases, the mean or average is not a very good representation of the results and comparisons between groups with large error bars may not be meaningful.

Figure 1 shows that there was not much difference in the overall number of errors between the PP and VRR systems (leftmost bars) (PP M = 15.4%, VRR M = 15.9%). Moderate difficulty scenarios had a few more errors, and departure problems had higher errors. Further analysis of the departure scenario data showed that 41 of the 70 VRR errors were in one run and 15 of these were due to the problem the VRR had in recognizing the number “four.” (We have not included any analyses of statistical significance between PP and VRR runs for any of the data sets due to insufficient data points.)

The nature of the errors was very different between the PP and VRR systems. The PP errors were generally inadequate readbacks (e.g., not including call sign), and failure to key in the clearances correctly or in time (and in some cases not keying them in at all), causing traffic deviations. The VRR errors, instead, were “say agains” and incorrect readbacks, which were generally caught by the controller and corrected. (The VRR system always executed the clearance that it read back, whereas the PP may have read back correctly, but sometimes made the wrong inputs.) The VRR system had fewer deviations, and those that did occur were often

due to software bugs that could be corrected. (Errors are expressed in terms of percent of total push-to-talks [PTTs].)

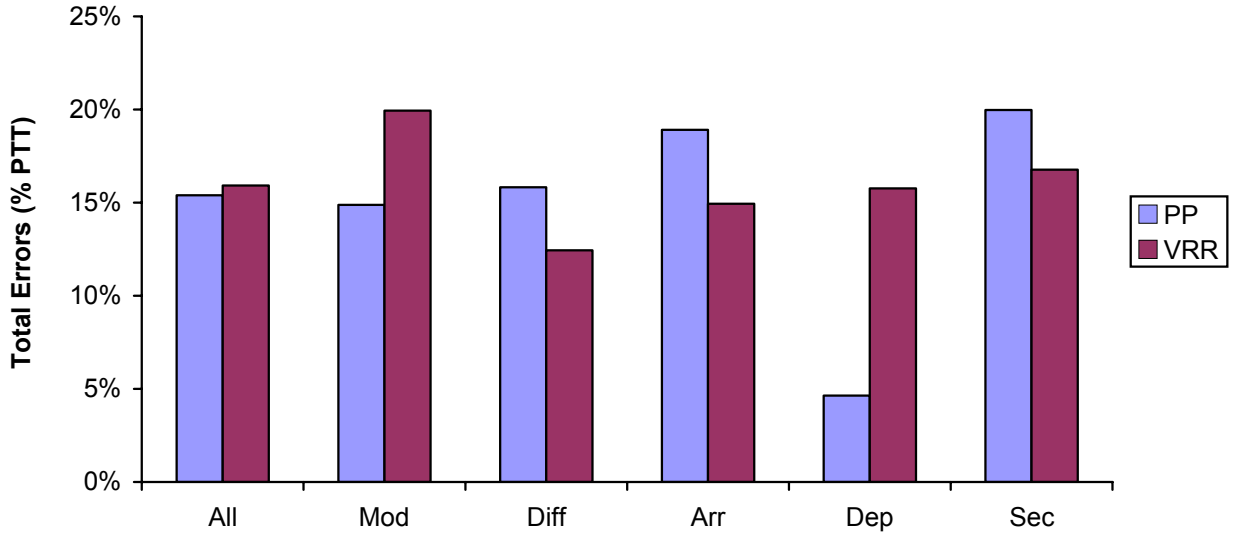

Figure 1. Total error rates for PP and VRR systems.

Figure 2 shows only the "consequential" errors for PP and VRR. These errors were those that created a significant problem in the simulation or disrupted the training, as judged by the instructor observing the runs. Although the differences are quite small, the VRR system performed somewhat better overall (PP M = 1.1%, VRR M = 0.4%), and better in the different types of scenarios. There were no consequential errors for either system in the departure scenarios.

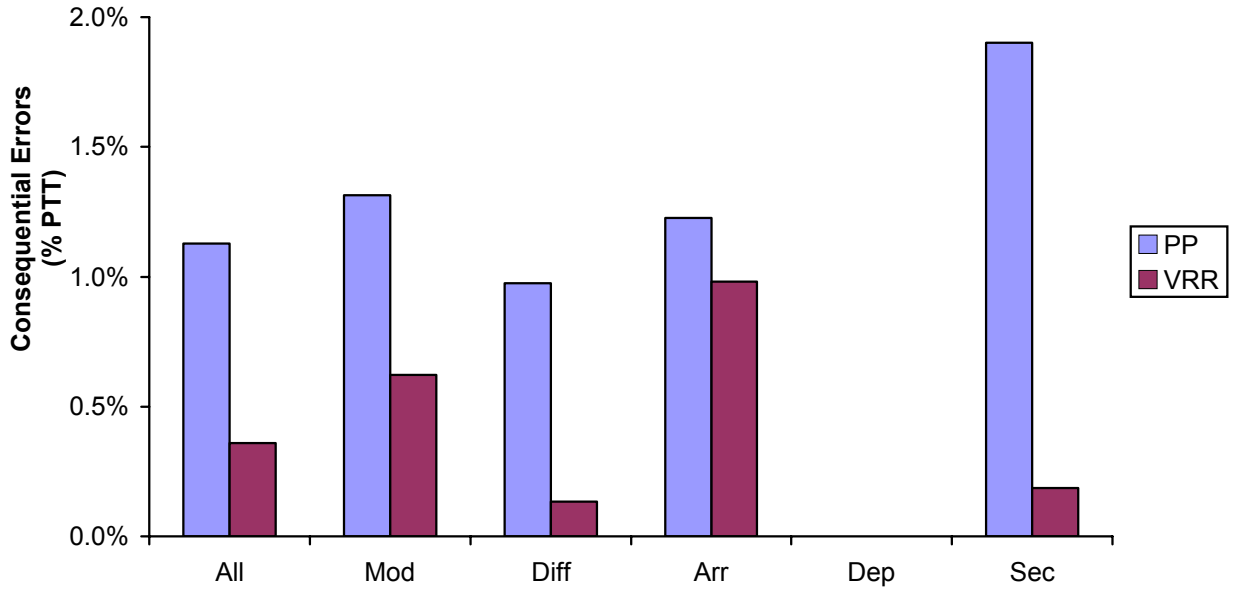

Figure 2. Consequential error rates for PP and VRR systems.

These graphs do not show, however, that the VRR system had specific difficulties with two of the controller participants. There were consistent problems with these controllers in that the system could not reliably recognize some of their clearances. As a result, two of the four runs

attempted by these controllers were terminated due to excessive VRR errors. There were no detectable issues with how these controllers enunciated words or expressed clearances that would suggest the nature of the problems.[1]

The results for the post run questionnaires completed by the instructor and controllers are organized by question.

"Please rate how well the VRR system or PP recognized controller instructions."

Figure 3 shows the results for this question. There was agreement between controller participants and the instructor in the ratings for the PP and VRR systems. The combined ratings for controllers and the instructor ("Both") showed the PP option to be nearly one point higher than VRR (PP $M = 4.2$, VRR $M = 3.3$). This suggests that the users found the human PP generally recognized their clearances better than the VRR system. However, any rating above 3.0 is in the "acceptable" range. (These data included those of the two participants where the system had problems. They gave very low ratings to VRR. The error bars also show that there was considerable variation in the VRR responses. This seems to mirror our observations that the VRR worked well with some controllers, and not so well with others.)

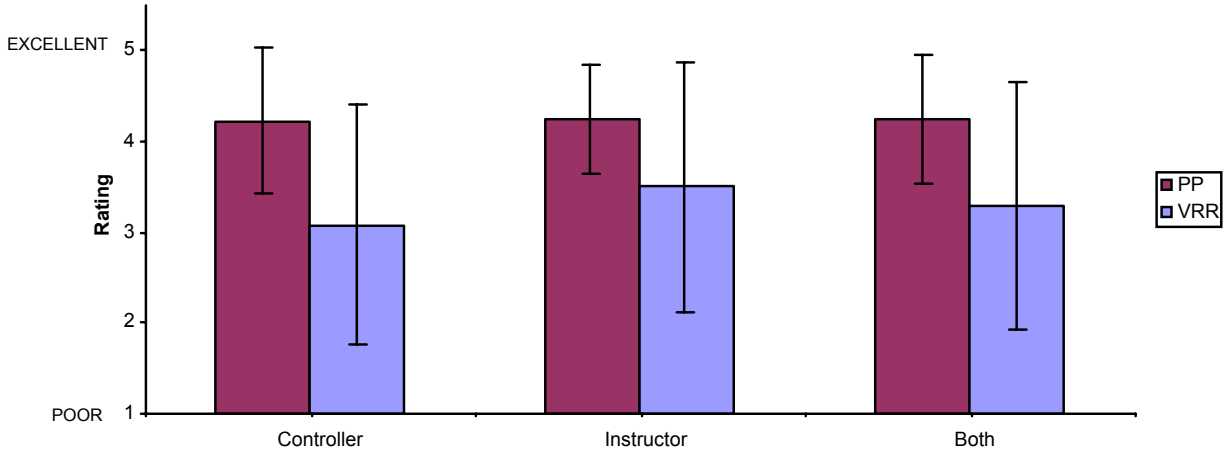

Figure 3. Average ratings for how well PP and VRR systems recognized controller instructions.

"If there were voice recognition problems, how disruptive were they?"

Figure 4 shows the results for this question. The controllers and instructor were not in agreement about the VRR system. The instructor seemed to feel the VRR problems were less disruptive than the controllers. The combined ratings suggest that the VRR system errors were rated as a little more disruptive than the PP alternative (PP $M = 1.6$, VRR $M = 2.3$), though the difference may not be meaningful. Opinions varied, perhaps due to different experiences with the VRR

[1] One might expect many consequential errors in these two runs. This was not the case. Consequential errors were generally those that resulted in operational errors, missing the localizer, "wandering" aircraft, and other disruptive events. The two runs that were terminated did not go on long enough to generate many consequential errors, but the quantity of missed recognitions made it clear that continuing was not feasible.

system. The average scores are below 3, or the midpoint, suggesting that the problems were not very disruptive, on average.[2]

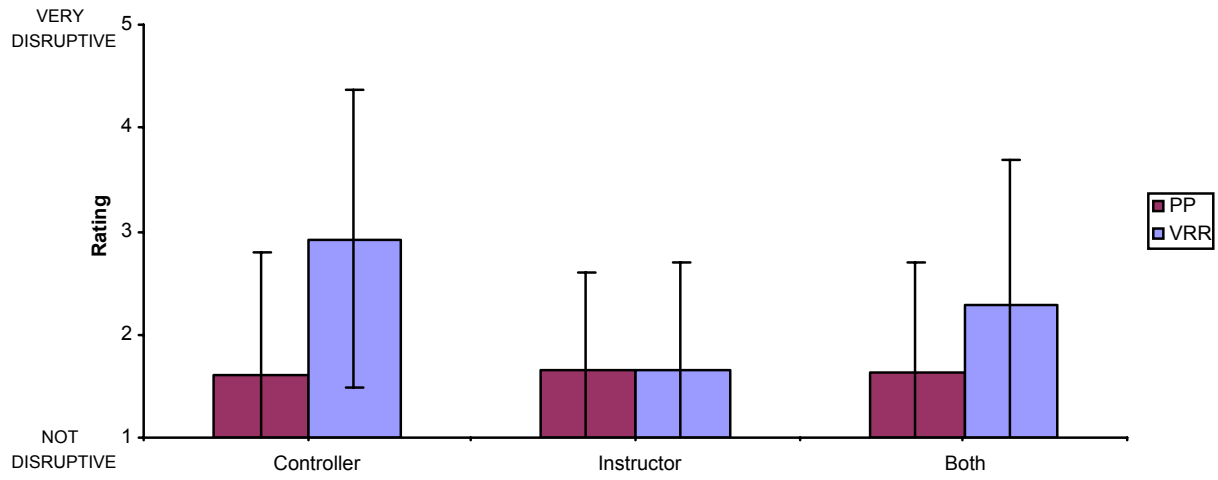

Figure 4. Average ratings for disruptiveness of PP and VRR voice recognition problems.

"Please rate how well you could understand the VRR or PP voice readbacks."

In this case, (see Figure 5) the instructor rated the quality of the VRR voices as slightly lower than the controllers, and also rated the PP voices as lower in intelligibility. The combined ratings for the two systems are very similar and the difference is not meaningful (PP $M = 4.3$, VRR $M = 4.1$). Both systems had acceptable voice output.

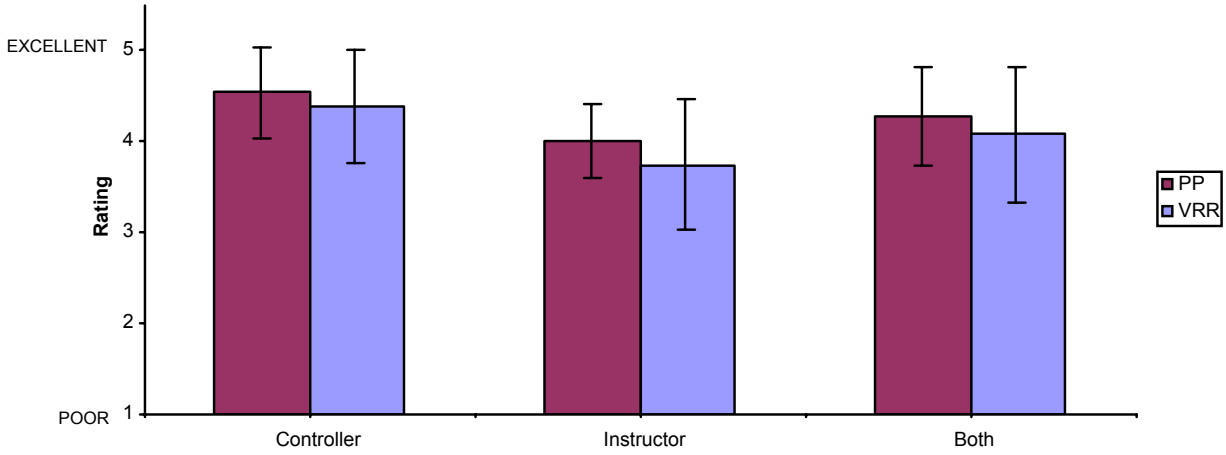

Figure 5. Average ratings for the quality of readbacks from PP and VRR systems.

[2] This might be a good example of how averaged data tend not to show specific problems, such as the need to stop two of the runs due to VRR errors. This is why several data sources need to be considered when evaluating any new system.

"If it was difficult to understand the VRR or PP's voice, how disruptive was this?"

Figure 6 shows that the instructor rated the effects of not understanding the VRR and PP voices nearly the same as the controllers. When combined, however, the results were the same, and low (PP $M = 1.1$, VRR $M = 1.1$), suggesting that the readbacks from either system were not problematic under any conditions.

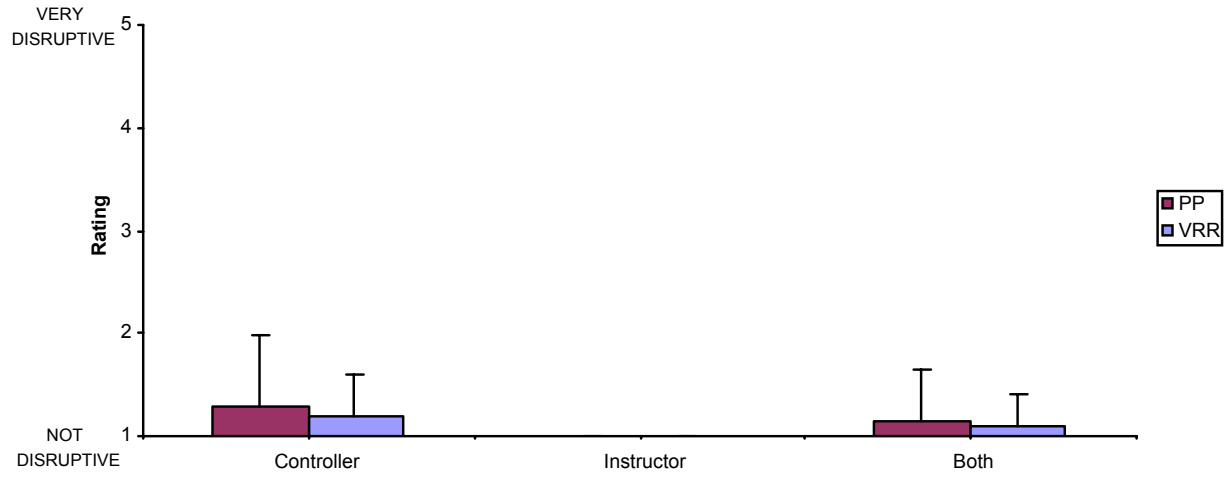

Figure 6. Average ratings for effects of not understanding PP or VRR voice output.

"Did the way in which the VRR or pseudo pilot operated cause the controller to change the way in which you worked in any way? If YES, how?"

For this question, the respondents answered either "yes" or "no" for whether using the PP or VRR systems changed their operational style. (The instructor was evaluating the controller, not commenting on his own operating style.) The results are shown in Table 2. For the PP system, the controllers most often said they did not change their style, whereas the instructor more frequently observed that the controller's style was affected. The instructor noted that controllers tended to work interactively with the PP, changing their style to assist as PP workload increased. This was not the case for the VRR runs.

Eight out of twelve controllers indicated that they changed their approach when using the VRR system, whereas the instructor's opinions of this were equally divided (sometimes did, and sometimes did not). The controllers made comments about how they felt the VRR system affected them. (Written comments for all questions can be reviewed in Appendix C.)

Table 2. Changes in operating style as a function of simulator type.

	Yes	No
Pseudo Pilot		
Controller	1	11
Instructor	8	4
Total	**9**	**15**
Voice Recognition and Response		
Controller	8	4
Instructor	6	6
Total	**14**	**10**

The controllers offered some useful comments during our discussions after the last test session of each day. Some said that they needed to work more slowly for the VRR system. They had to wait for full (lengthy) readbacks whereas in the real world readbacks are often abbreviated as workload increases. The PP also tends to shorten readbacks when busy. The VRR system made the controllers pay more attention to readbacks since there were sometimes errors, whereas the PP readbacks were usually correct. Some controllers noted they had to be more careful about how they spoke for the VRR system and thought that this might be distracting for a developmental. A good summary might be, as said by one participant: "Neat, but needs some work."

3.3 Phase 1 Discussion

The VRR system did not require significantly more time to prepare or build scenarios as compared to the existing PP system. Resources were needed, however, for improvements, such as adding to the vocabulary files.

Generally, the VRR system seemed to perform comparably to the PP system, although in a few cases it performed worse (because of voice recognition problems with specific controllers). It was observed to work equally as well as the PP in several test runs (particularly the complex high traffic cases) and sometimes even performed better than the PP. The VRR system really seemed to come into its own in complex, high demand situations.

The error counts do not indicate large differences between the PP and VRR systems. It was very useful to collect the PP error data since it showed us the types and frequency of errors that a skilled PP could make running moderate and difficult scenarios. Many of the PP and VRR errors are tolerable since the controller can correct them. However, some are disruptive in that they create distractions, operational errors, or other problems that should not occur during training. It is in looking at the consequential error data that it appears that the VRR system is superior to the PP.

One problem with the PP system is that, while the PP might give a correct readback, this does not mean that the data entry into the simulation system is accurate. This sometimes resulted in disruptive aircraft deviation errors that were only caught by the controller some time later in the training run. This could prove very distracting to developmentals. The VRR system always acts in accordance with its readback, even though this may be incorrect at times.

Nevertheless, the VRR system did have acute voice recognition problems with two controllers where the simulation sessions had to be terminated. The voice engine did not seem to be able to recognize standard clearances issued by these controllers. There were other instances of poor recognition in other runs, but we were able to continue with training to completion. In some cases, it was possible to make changes to the vocabulary files and program to fix VRR problems. However, improving VRR responses for these speakers required system changes.

Although opinions differed, questionnaire results showed that controllers and instructors thought that the speech recognition function of the VRR system was not quite as good as compared to a human PP. Any problems resulting from this were not very disruptive. Readback quality was very good and, when there were issues, they were not disruptive. Controllers tended to change their operating style to adjust to the VRR system more often, as compared to the PP.

4.0 PHASE 2

4.1 Method

4.1.1 Equipment

The same equipment was used as in Phase 1. Improvements were made to the voice recognition engine, vocabulary files, and voice output after the Phase 1 data collection.

4.1.2 Participants

The participants were six CPCs from Boston Consolidated TRACON. Three controllers had participated in the Phase 1 study. The VRR system had difficulties with one of these controllers in Phase 1 to the point where the run had to be terminated. The second Phase 1 controller where the VRR system had serious problems was not available during the week of Phase 2 data collection.

4.1.3 Scenarios

The scenarios were identical to those run in Phase 1.

4.1.4 Experimental Design

The approach was the same as in Phase 1, except that we did not gather any further data from PPs.

4.1.5 Data Collection

Data collection procedures were the same as in Phase 1. Phase 2 focused on checking for improvements in the VRR system. Only VRR scenarios were run in order to compare system performance with Phase 1 VRR runs. The schedule of runs for Phase 2 is found in Appendix A and was identical to Phase 1.

4.2 Results

The results for Phase 2 are based on the 12 data runs collected from the six participating controllers. In two cases, the VRR system had significant problems and the runs had to be terminated. Troubleshooting the system suggested some technical improvements that could be made immediately (e.g., to the vocabulary file). It was also observed that several controllers were using non-standard phraseology that caused difficulties for the VRR system. VRR only recognizes those words and phrases already in its vocabulary file. System improvements were made and the participants were counseled to use standard FAA phraseology. The identical problems were run a second time much more successfully. The two problematic runs were excluded from the data set.

Figure 7 shows a graph of the total number of errors for Phase 1 PP runs and Phases 1 and 2 VRR runs. This includes all of the error types shown in Table 1. Plots are included for moderate and difficult runs, and for the different types of scenarios (arrival, departure, and sector).

The results show that the average number of VRR errors across scenario types were about the same or lower than in Phase 1. Overall, Phase 2 VRR errors were lower (All: PP $M = 15.4\%$, VRR P1 $M = 15.9\%$, VRR P2 $M = 11.3\%$). This was mostly accounted for by fewer errors in moderate traffic load problems, and in arrival and departure sectors. The errors stayed about the same in difficult problems and sector scenarios.

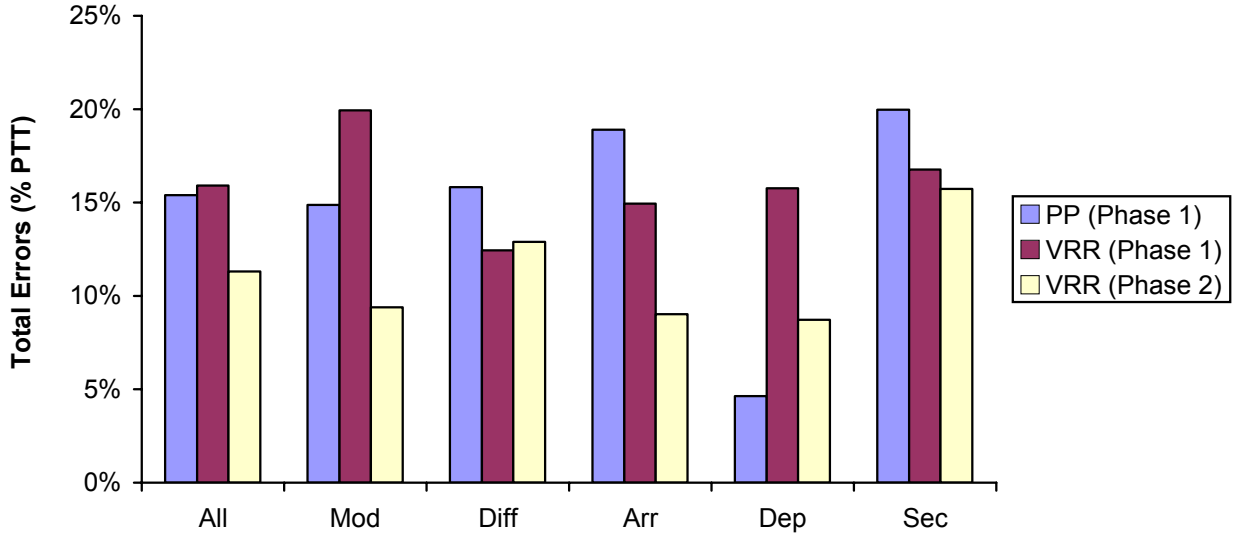

Figure 7. Total error rates for PP and VRR systems.

Figure 8 shows only the consequential errors for Phase 1 PP and Phases 1 and 2 VRR. The overall number of these errors went down for Phase 2 (All: PP M = 1.1%, VRR P1 M = 0.4%, VRR P2 M = 0.2%). However, it should be noted that the number of these errors was quite low in VRR runs (only five in Phase 1 and three in Phase 2) and therefore this difference, by itself, should not be considered as conclusive. The reduction seems to have been in moderate scenarios and arrival problems. There was a small increase in errors for difficult and sector problems.

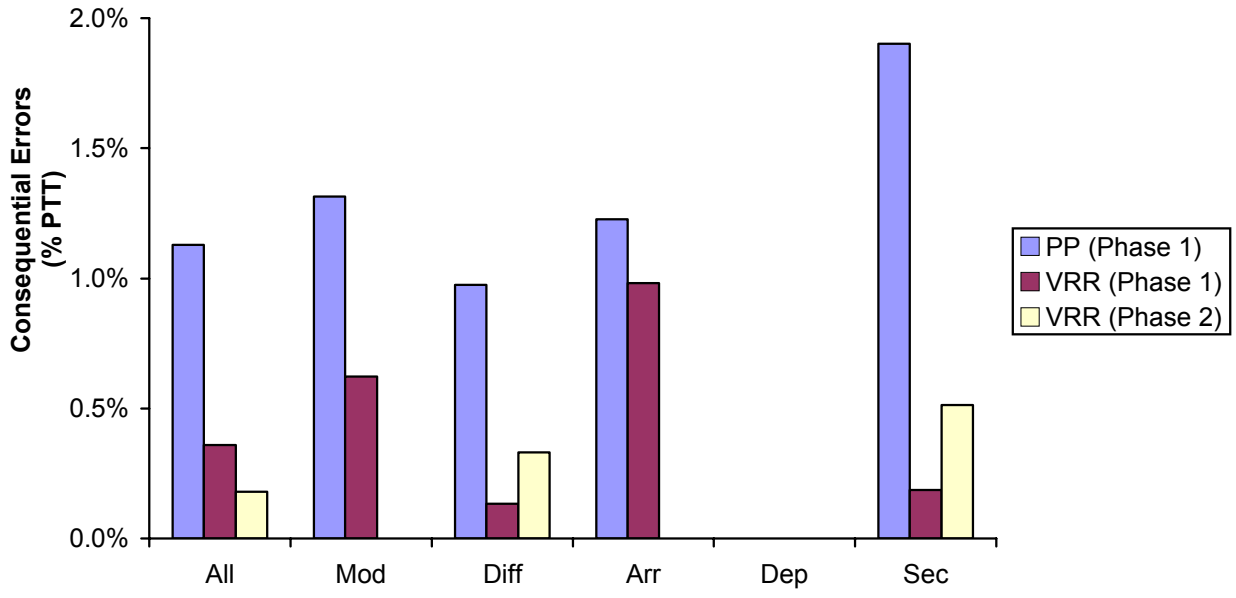

Figure 8. Consequential error rates for PP and VRR systems.

The results for the Phase 2 post run questionnaires completed by the instructor and controllers are presented by question.

"Please rate how well the VRR system or PP recognized controller instructions."

Figure 9 shows the results for this question. The ratings by both the instructor and controllers increased somewhat compared to Phase 1. This was reflected in the overall scores (PP $M = 4.2$, VRR P1 $M = 3.3$, VRR P2 $M = 3.7$). This suggests that the users found the VRR system to be a little better at recognizing their inputs in Phase 2. (There were four more responses at "acceptable" or higher than in Phase 1 out of a total of 24.) The error bars show more uniformity of opinion for VRR in Phase 2. All ratings are at or above 3 ("acceptable").

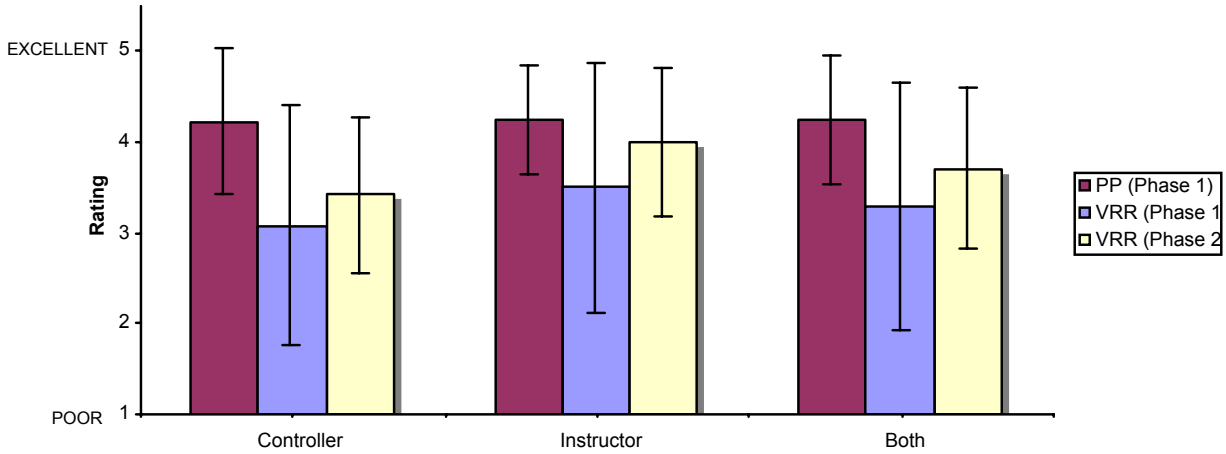

Figure 9. Average ratings for how well PP and VRR systems recognized controller instructions.

"If there were voice recognition problems, how disruptive were they?"

Figure 10 shows the results for this question. The ratings for VRR in Phase 2 were about the same and suggests that errors may be a little more problematic than the PP system (PP $M = 1.6$, VRR P1 $M = 2.3$, VRR P2 $M = 2.2$). There is a lot of variation in the responses, and indicates the average differences between PP and VRR are probably not be meaningful.

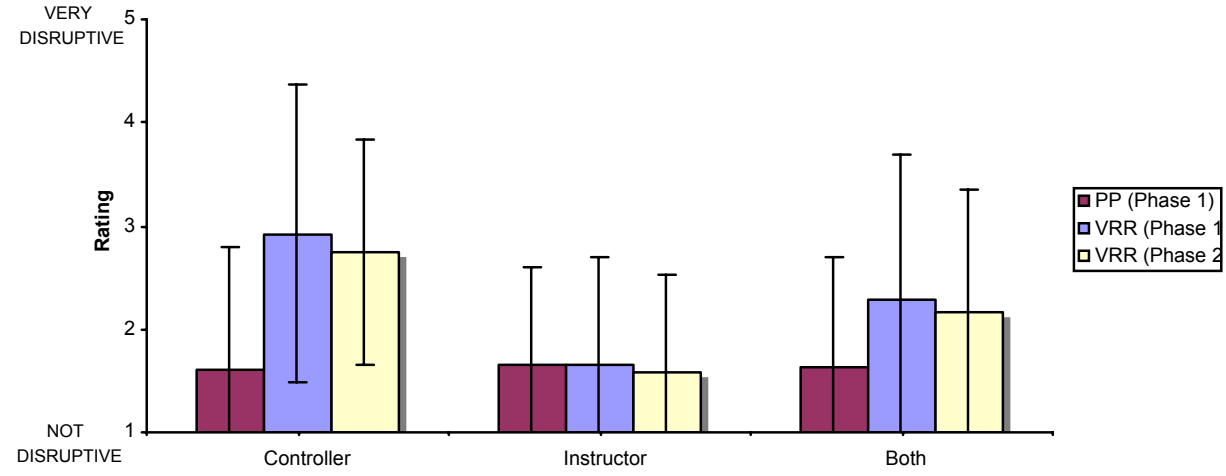

Figure 10. Average ratings for disruptiveness of PP and VRR voice recognition problems.

"Please rate how well you could understand the VRR or PP voice readbacks."

In this case, (see Figure 11) the instructor's ratings of the quality of the VRR voices were somewhat higher than in Phase 1, and a little better than the PP ratings in Phase 1. The controllers' ratings did not change. (We should recall that there were no PP runs in Phase 2 to compare to.) The average ratings for the two systems are very similar and the difference is probably not meaningful (PP $M = 4.3$, VRR P1 $M = 4.1$, VRR P2 $M = 4.4$). All ratings are in the very good to excellent range.

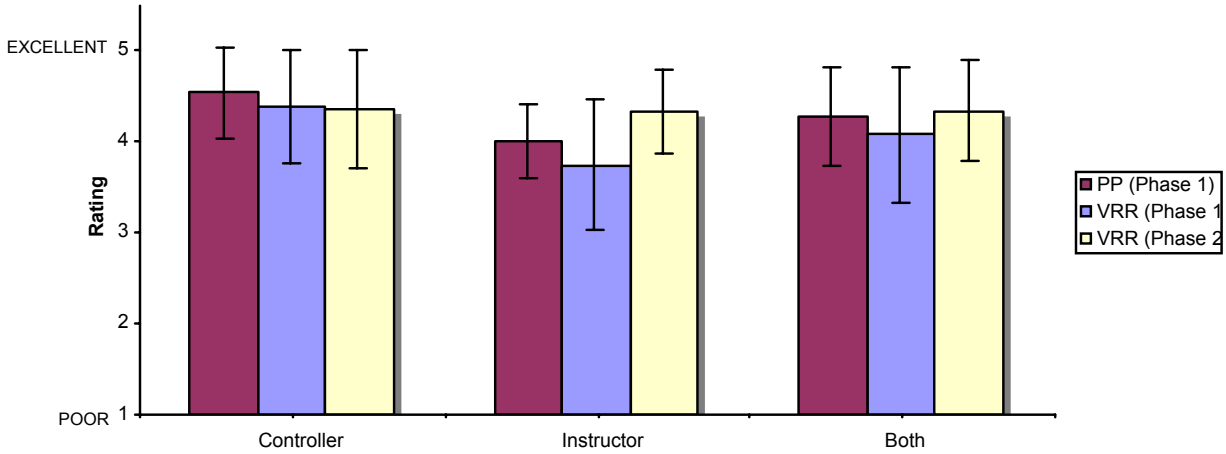

Figure 11. Average ratings for the quality of readbacks from PP and VRR systems.

"If it was difficult to understand the VRR or PP's voice, how disruptive was this?"

Figure 12 shows that the ratings for this question were slightly higher than in Phase 1, though there was considerable difference of opinion. The averages were: PP $M = 1.1$, VRR P1 $M = 1.1$, and VRR P2 $M = 1.6$. However, all ratings were still well below 2 out of 5, indicating that the level of disruptions was acceptable. The slight increase appears to be due to two runs that received negative ratings.

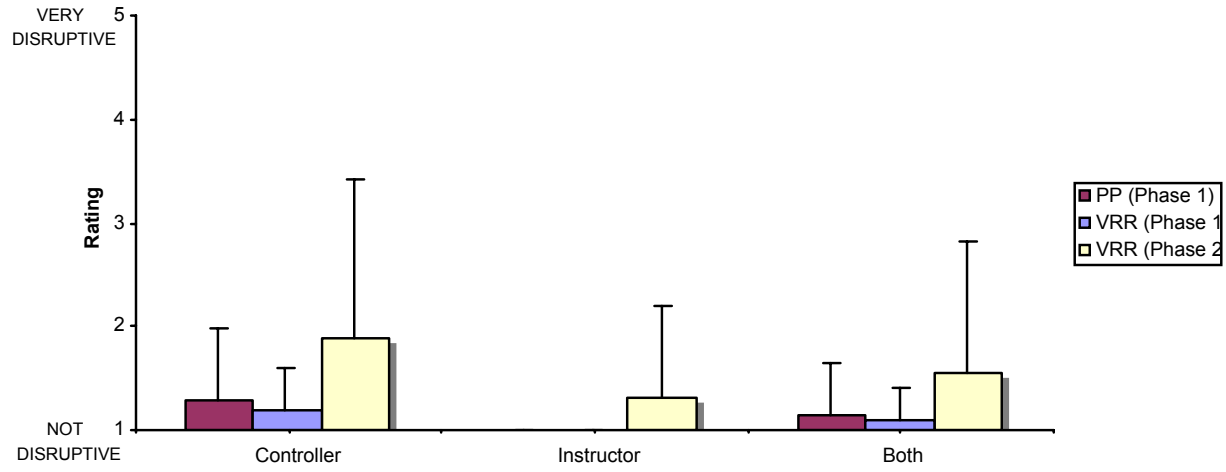

Figure 12. Average ratings for effects of not understanding PP or VRR voice output.

"Did the way in which the VRR or pseudo pilot operated cause the controller to change the way in which you worked in any way? If YES, how?"

Table 3 suggests that the controllers thought that using the VRR system caused them to change their control style a little less than in Phase 1. However, the difference is so small that it is probably not meaningful. The instructor's response showed a shift toward fewer changes in controller style in Phase 2. There were several comments that suggest that the controllers needed to adopt a slower speech rate and limit the length of multiple clearances for VRR as compared to PP. Written comments for all questions can be reviewed in Appendix D.

Table 3. Changes in operating style using VRR in Phases 1 and 2.

	Yes	No
Phase 1		
Controller	8	4
Instructor	6	6
Total	**14**	**10**
Phase 2		
Controller	7	5
Instructor	5	7
Total	**12**	**12**

4.3 Phase 2 Discussion

In general, the performance of the VRR system improved as compared to Phase 1. It was able to successfully negotiate the speech of one of the controllers where it previously had significant problems. Where difficulties did emerge, quickly changing some system parameters or suggesting modifications to participants' phraseology made a great deal of difference in system performance.

In Phase 2, the nature of the errors also shifted. In Phase 1 the errors were dominated by recognition mistakes (such as "four" and "niner") when controllers used standard FAA phraseology. These were sometimes severe enough to force termination of a run. In Phase 2, these types of errors were greatly diminished. When they appeared many were quickly fixed by changing certain VRR parameters. In Phase 2, several error types were observed:

a. Errors related to unsupported simulator functionality, such as "expedite." Rectifying these issues will require adding new capabilities to the AT Coach simulator, as opposed to the VRR system.

b. There were errors due to controller use of inappropriate phraseology that the VRR system did not have in its grammar set. Some of these were quickly corrected by adding words to the vocabulary dictionary. In some cases, however, alerting the controller to the appropriate phraseology improved performance considerably. Particular instances of non-standard phraseology included truncated terms (e.g., "Alaska" instead of "Air Alaska"), missing words (e.g., failure to use "correction" to recover from stumbled clearances), or poor pronunciation (e.g. "Novemmer" instead of "November").

c. The VRR had particular problems with complex or multi-part clearances, which, even though sufficiently correct from a phraseology standpoint, caused recognition errors.

d. Delay was observed in the VRR readback responses. This sometimes caused a repetition of the command by the controller. This delay might be due to the extra processing required as a result of increasing capabilities to eliminate recognition errors (e.g., using "gender biasing").

e. The VRR was observed to have a deficiency in its automatic voice input level control functions. For one controller, the VRR clipped the upper and lower boundaries of the voice envelope. This resulted in poor recognition performance for that controller. Asking him to speak more softly and move his microphone helped improve recognition.

f. Errors sometimes occurred when alerting aircraft to other traffic. In these cases the phraseology used was inconsistent among controllers and resulted in VRR errors.

g. The VRR responded with "runway not found" to some controller instructions. Rather than responding with "say again," for example. This error was not previously heard in Phase 1.

While it was necessary to terminate two runs in Phase 2 due to recognition problems, corrections were quickly made that allowed a re-start and successful completion of these scenarios.

In general, questionnaire ratings in Phase 2 showed some improvement for voice recognition over Phase 1, with other responses being similar to Phase 1.

5.0 CONCLUSIONS FROM PHASES 1 AND 2

Scenario preparation times are similar between the PP and VRR systems. There is no substantial difference in the time required to set up and run VRR training scenarios as oppose to PP runs.

In Phase 1, VRR performed well as compared to the existing PP system. During the 10 assessment runs that were fully completed, the types of errors tended to result in fewer disruptions to training than those in PP runs. However, some problems existed with voice recognition for specific users that needed to be rectified.

In Phase 2, there were fewer errors. The system did much better with voice recognition for users who had difficulty in Phase 1. Also, the type of errors shifted away from recognition of words that might be found in standard FAA phraseology to more specific, well-defined problems that could be rectified on the spot, or dealt with through future software changes. It also became evident that improvements could be made simply by asking controllers to change their phraseology over to more standard forms. In some cases, it may not be desirable in a training system to accommodate all controller styles, when developmentals are required to use standard words and phrases.

From our experience with the Phase 1 and 2 evaluations, it is clear that the assessment exercise has been valuable for VRR system development. The results of Phase 1 were used to make improvements that resulted in better system performance in Phase 2. In the next few months, developmentals will start training at Boston Consolidated TRACON. We expect that the system will perform better with this group than with experienced controllers. Developmentals will tend to use standard phraseology and the scenarios they use will be less difficult than those we tested.

We recommend that the VRR developer continues with improvements to the VRR system to enhance its accuracy, based on the findings of this evaluation. We also recommend a Phase 3 test of the VRR system with developmentals. This is important since the VRR capability will frequently be used for developmental training and it has not yet been assessed with this group. VRR should also work adequately well with most experienced controllers for training purposes. However, efforts will be needed with some participants to help them use standard phraseology while using this system.

Assuming that improvements are made based on our findings, the system should be ready to deploy to other sites. Testing with developmentals could be completed at Boston Consolidated TRACON, or could be done at the next deployment location. It might be best to field the system in stages, and assess its success at a limited number of other facilities before proceeding further.

When the VRR system is installed at other sites, there will no doubt be a need to make local adjustments and configuration changes. It will be important for on-site staff to work with the system and have the support of its developers to make adjustments, as needed.

6.0 ACKNOWLEDGEMENTS

The author appreciates the assistance provided by Tom Norato (FAA), Coleman Hartigan (FAA), Bill Preston (QSS), Tony Evans (L-3 Titan), Felipe Cuevas (San Jose State University), Husni Idris (L-3 Titan), and Steve DePascale (L-3 Titan).

APPENDIX A

SCHEDULE

KEY FOR TABLES

Difficulty/Complexity
Moderate = 90%
Difficult = 110%

Scenario Types and Runway Configurations
D1 = Departure, 4R/L-9
D2 = Departure, 33L/27
A1 = Arrival, 4R/L-9
A2 = Arrival, 27/27
S1 = Sector, 33L-27 Plymouth
S2 = Sector, 4R/L-9 Rockport

PHASE ONE

Run	System	Controller	Pseudopilot	Difficulty	Arr/Dep/Sec	Config
1	VRR	A		Moderate	A	1
2	PP	B	A	Moderate	D	1
3	VRR	B		Moderate	D	1
4	PP	A	A	Moderate	A	1
5	VRR	A		Difficult	D	2
6	PP	B	B	Difficult	S	2
7	VRR	B		Difficult	S	2
8	PP	A	B	Difficult	D	2
9	VRR	C		Moderate	A	2
10	PP	D	B	Moderate	D	2
11	VRR	D		Moderate	D	2
12	PP	C	A	Moderate	A	2
13	VRR	C		Difficult	D	1
14	PP	D	A	Difficult	S	1
15	VRR	D		Difficult	S	1
16	PP	C	A	Difficult	D	1
17	VRR	E		Moderate	S	1
18	PP	F	B	Moderate	S	2
19	VRR	F		Moderate	S	2
20	PP	E	B	Moderate	S	1
21	VRR	E		Difficult	A	1
22	PP	F	B	Difficult	A	2
23	VRR	F		Difficult	A	2
24	PP	E	B	Difficult	A	1

PHASE TWO

Run	System	Controller	Difficulty	Arr/Dep/Sec	Config
1	VRR	1	Moderate	A	1
3	VRR	2	Moderate	D	1
5	VRR	3	Difficult	D	2
7	VRR	4	Difficult	S	2
9	VRR	5	Moderate	A	2
11	VRR	6	Moderate	D	2
13	VRR	7	Difficult	D	1
15	VRR	8	Difficult	S	1
17	VRR	9	Moderate	S	1
19	VRR	10	Moderate	S	2
21	VRR	11	Difficult	A	1
23	VRR	12	Difficult	A	2

APPENDIX B
DATA COLLECTION FORMS

Date	
DP Number	
Start Time (Z)	
Total PTTs	

Error Codes
N = No immediate readback
R = Readback incorrect
P = Pilot says "Say again"
C = Controller says "Say again"
D = Pilot deviation
Consequences: Y = Yes; N = No

Error Code	ACID	Time (Z)	Notes	Conseq.
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N

Date	
DP Number	
Start Time (Z)	
Total PTTs	

Error Codes
N = No entry
I = Incorrect entry (fixed or not?)

Error Code	ACID	Time (Z)	Notes	Conseq.
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N
				Y N

Voice Recognition and Response (VRR) Controller Feedback

Controller: _____ (A – M) **Date: _____** **DP Number: _____**

1. Please rate how well the VRR system or pseudo pilot recognized your instructions.

1	2	3	4	5
Poor		Acceptable		Excellent

Comments:

_____

_____

_____

_____

2. If there were voice recognition problems, how disruptive were they?

1	2	3	4	5
Not Disruptive		Somewhat Disruptive		Very Disruptive

Comments:

_____

_____

_____

_____

3. Please rate how well you could understand the VRR or pseudo pilot voice readbacks.

1	2	3	4	5
Poor		Acceptable		Excellent

Comments:

_____

_____

_____

_____

4. If it was difficult to understand the VRR or pseudo pilot's voice, how disruptive was this?

1	2	3	4	5
Not Disruptive		Somewhat Disruptive		Very Disruptive

Comments:

_____

_____

_____

_____

5. Did the way in which the VRR or pseudo pilot operated cause you to change the way in which you worked in any way?

YES NO

If YES, how?

_____

_____

_____

_____

6. Do you have any comments on the VRR system?

_____

_____

_____

_____

Voice Recognition and Response (VRR) Instructor Feedback

Controller: _____ (A – M) Instructor: _____

Date: _____ DP Number: _____

1. Please rate how well the VRR system or pseudo pilot recognized the controller's instructions.

1	2	3	4	5
Poor		Acceptable		Excellent

Comments:

_____

_____

_____

_____

2. If there were voice recognition problems, how disruptive were they?

1	2	3	4	5	N/A
Not Disruptive		Somewhat Disruptive		Very Disruptive	

Comments:

_____

_____

_____

_____

3. Please rate how well could you understand the VRR or pseudo pilot voice readbacks.

1	2	3	4	5
Poor		Acceptable		Excellent

Comments:

_____

_____

_____

_____

4. If it was difficult to understand the VRR or pseudo pilot's voice, how disruptive was this?

1	2	3	4	5	N/A
Not Disruptive		Somewhat Disruptive		Very Disruptive	

Comments:

_____

_____

_____

_____

5. Did the way in which the VRR or pseudo pilot operated cause the controller to change the way in which he/she worked in any way?

YES NO

If YES, how?

_____

_____

_____

_____

6. Do you have any comments on the VRR system?

_____

_____

_____

_____

APPENDIX C
PHASE 1 WRITTEN COMMENTS

Phase 1 Controller Comments (By Question)

Voice Recognition and Response

*Please rate how well the **VRR** system recognized instructions.*

Run Number	Comment
DP1	Need I say more? [Refers to assigned rating of "Poor."]
DP5	Would not recognize "4." Took a turn instead of frequency. No response on a few.
DP7	A few missed calls.
DP11	KAP/UCA - numerous mix-ups.
DP19	VRR did not recognize "niner" in altitude or altimeter.
DP23	Misunderstood some nines and fives.

*If there were **VRR** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP2	No problems.
DP4	There were no voice recognition errors.
DP24	They never happened at a critical time. Simulated real life.

*Please rate how well you could understand the **VRR** readbacks.*

Run Number	Comment
DP9	They were readable.
DP11	It almost speaks too slow.
DP13	The readbacks were loud and clear for the most part. Sometimes we would not get any response.
DP19	UCA call sign a little hard to understand at first.

*If it was difficult to understand the **VRR**, how disruptive was this?*

Run Number	Comment
DP9	Just different listening for computer voice.
DP21	Just getting used to the read back of the clearance.

*Did the way in which the **VRR** operated cause you to change the way in which you worked in any way? If YES, how?*

Run Number	Comment
DP1	In the future it wouldn't, just getting used to the system and what it can and can't do.
DP9	Nobody was complying w/instructions.
DP11	I had to be very careful in my pronunciation. I had to think about whether or not the VRR was going to understand which diverted my attention from my job. Training would have been useless.
DP13	I did not feel as though I could give instructions in rapid succession like I would need to on the floor.

DP15	I couldn't do my job.
DP17	I slowed my speech rate when I got a lot of 'say again' readbacks.
DP19	Slowed speech rate somewhat.
DP21	Yes - asking if traffic in sight and then the A/P in sight.

Pseudo Pilot

*Please rate how well the **PP** recognized instructions.*

Run Number	Comment
DP4	Some a/c wouldn't acknowledge transmissions.
DP18	Aircraft never acknowledged instruction or they acknowledged instruction and did not comply.
DP20	No problems w/readbacks; much better than reality.
DP22	This problem ran a lot smoother although there were a few bad readbacks.
DP24	A couple of times the aircraft went through the localizer.

*If there were **PP** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP2	No problems
DP4	There were no voice recognition errors.
DP18	N/A
DP22	N/A
DP24	They never happened at a critical time. Simulated real life.

*Please rate how well you could understand the **PP** readbacks.*

Run Number	Comment
DP12	No problems
DP14	A few deleted items in readbacks.

*If it was difficult to understand the **PP**, how disruptive was this?*

Run Number	Comment
DP4	Not too difficult to understand
DP14	But this is normal ATC ops.

*Did the way in which the **PP** operated cause you to change the way in which you worked in any way? If YES, how?*

Run Number	Comment
DP12	I would wait until I thought the pilot was caught up before giving another instruction.

Phase 1 Instructor Comments (By Question)

Voice Recognition and Response

*Please rate how well the **VRR** system recognized instructions.*

Run Number	Comment
DP5	Failed to recognize the number "four". Many misrecognitions.
DP1	Marginal improvement over previous.
DP15	Overall, pretty good - some problems w/"14" thousand and barging.
DP19	Problem w/"niner".

*If there were **VRR** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP9	Failure to recognize/execute commands.
DP15	Required constant correction.
DP17	Minor grammar changes needed.
DP23	"Niner" needs work.

*Please rate how well you could understand the **VRR** readbacks.*

Run Number	Comment
DP3	Some of the voice fonts are in need of tweaking
DP1	Voice fonts need tweaking
DP23	Voice fonts need work.

*If it was difficult to understand the **VRR**, how disruptive was this?*

Run Number	Comment
	(No comments.)

*Did the way in which the **VRR** operated cause you to change the way in which you worked in any way? If YES, how?*

Run Number	Comment
DP11	More attentive to readbacks particularly in light of the # of misreads-backs. Pacing of pilots readbacks/init call-ups.
DP13	The controller was hesitant, didn't "trust" the system could keep up. The controller did not have their display set up, as normal.
DP15	Anticipated errors to occur/require amendments.
DP17	Couldn't use "14" thousand consistently. Some grammar changes necessary.
DP19	Seem to keep pace equivalent to pseudo pilot run.
DP21	Minor grammar changes required.
DP23	Improvement over previous day: 1) Still need abbreviated clearances to "flow" better, 2) Voice fonts need work.

Pseudo Pilot

*Please rate how well the **PP** recognized instructions.*

Run Number	Comment
DP6	Some minor grammar issues, but overall very well.
DP18	Failed to recognize instructions to: 1) intercept the localizer, 2) headings

*If there were **PP** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP18	Many readback errors.
DP22	Many "say agains" or incorrect/hesitant readbacks.

*Please rate how well you could understand the **PP** readbacks.*

Run Number	Comment
DP10	Felt that the pseudo pilot was slow in response.

*If it was difficult to understand the **PP**, how disruptive was this?*

Run Number	Comment
	(No comments.)

*Did the way in which the **PP** operated cause you to change the way in which you worked in any way? If YES, how?*

Run Number	Comment
DP6	They paced off each other and "co-managed" the scenario.
DP8	Several opps… then the controller would wait for the pilot to recall.
DP12	The controller seems to wait until the pseudo pilot was "caught up" before continuing.
DP14	Pseudo pilot was not in "pace" w/scenario resulting in increased workload.
DP16	Controller would anticipate the pseudo pilot readback/callup or "back-off" if a barge occurred.
DP18	Many readbacks, failure to use call signs in readback, many "say agains."
DP20	Pseudo pilot speed was slow and didn't pace well w/scenario.
DP22	No - controller handled the readback errors properly.
DP24	Pilot had difficulty keeping in pace. Considerable hesitation for readbacks and delays in command entries.

APPENDIX D
PHASE 2 WRITTEN COMMENTS

Phase 2 Controller Comments (By Question)

Voice Recognition and Response

*Please rate how well the **VRR** system recognized instructions.*

Run Number	Comment
DP1	There were a few times VRR did not recognize a runway (4R).
DP5	I reduced my speech rate from my normal. There were 1 or 2 instances where VRR did not recognize my voice. (Due to combined [transmissions].)
DP7	For the amount of traffic, the times where there was some confusion seemed realistic. It appeared some a/c took instructions for other a/c when I transposed the call sign.
DP9	Saying "roger" caused "ident."
DP11	Few mistakes compared to past scenarios.
DP13	Good recognition, a couple of times a/c read back random instructions.
DP15	Numerous errors
DP17	A few had trouble understanding traffic calls. Tried 3X before understanding.
DP19	VRR seemed to confuse fly heading 070 with Tr 070.
DP21	Pilot's weren't recognizing all [approach] clearances - "loop effect" if an a/c checks in and I stepped on him.
DP23	VRR confused 5 and "niner" often.

*If there were **VRR** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP1	This slight disruption was just a bump in the road compared to the last problem.
DP3	There were times when multiple departures were trying to key in right after I ended my transmission. But this actually is realistic.
DP5	As stated above, there was 1 or 2 times where I had to split my [transmissions] and move on.
DP7	It required additional transmissions so it increased the workload. I felt at times a complex or multi-instruction transmission would not be understood so I would break it up.
DP9	When pseudopilot did not recognize initial transmission they kept checking in.
DP11	Any repeated transmissions for a trainee would be disruptive.
DP15	Didn't do half what I normally would.
DP17	Making a [transmission] 3 times takes a lot of time away from other duties.
DP19	Needed to repeat some headings with arrivals over PUD to BOS.

*Please rate how well you could understand the **VRR** readbacks.*

Run Number	Comment
DP3	The pronunciation of some call signs were not quite correct. CHQ, KAP.
DP17	Julia's voice is too high for me. The rest are good.
DP21	Julia's voice - 1 - I don't care for the pitch - requires too much concentration to understand her

*If it was difficult to understand the **VRR**, how disruptive was this?*

Run Number	Comment
DP17	I just had to concentrate more on J's.

*Did the way in which the **VRR** operated cause you to change the way in which you worked in any way? If YES, how?*

Run Number	Comment
DP1	I felt that I slowed down my pace a bit.
DP5	I did slow down my speech rate. I felt that I was trying to accommodate VRR.
DP7	Again, control instructions that were two- or three-fold seemed to not be understood as well as simpler transmission, so this made me want to change the way I issued them.
DP9	Tried to use standard phraseology and speak slower.
DP13	Read frequencies clearer with the point emphasized.
DP15	If I can't say things like via 0.65, then what's the point. Half of what I would normally say, i.e., "FH 080 vectors ILS Rwy. 33L FA covrs."
DP23	Runway 4R + 4L was more difficult since transmissions had to be separated (maintain visual - cleared for approach).

Phase 2 Instructor Comments (By Question)

Voice Recognition and Response

*Please rate how well the **VRR** system recognized instructions.*

Run Number	Comment
DP1	Still needed (controller) prompting for use of standard phraseology.
DP11	Much improved use of "four."
DP19	1st 3 minutes were difficult for John's recognition.
DP23	Some difficulty w/"niner" vs. "five" readback.

*If there were **VRR** voice recognition problems, how disruptive were they?*

Run Number	Comment
DP21	Barging during readbacks.

*Please rate how well you could understand the **VRR** readbacks.*

Run Number	Comment
DP5	N/A

*If it was difficult to understand the **VRR**, how disruptive was this?*

Run Number	Comment
	(No comments.)

*Did the way in which the **VRR** operated cause you to change the way in which the controller worked in any way? If YES, how?*

Run Number	Comment
DP1	Standard phraseology for traffic calls needed to be prompted.
DP5	Yes. Non-standard phraseology would error in multi-command instructions.
DP9	Forced standard phraseology.
DP13	Forces good phraseology.
DP15	Numerous recognition errors that challenged the participant. In turn, the participant not engaged to the scenario. Some instances of non-standard phraseology would error.