

This paper was cleared by ASC-96-1774 on 10 Jul 1996.

FLIGHT TEST RESULTS OF ITT VRS-1290 IN NASA OV-10

David T. Williamson  
Pilot-Vehicle Interface Branch (WL/FIGP)  
WPAFB, OH  
(513) 255-6696  
williadt@wl.wpafb.af.mil

Timothy P. Barry  
North Coast Simulation  
Dayton, OH  
(513) 255-0240  
barrytp@wl.wpafb.af.mil

Kristen K. Liggett  
Pilot-Vehicle Interface Branch (WL/FIGP)  
WPAFB, OH  
(513) 255-8265  
liggetkk@wl.wpafb.af.mil

## ABSTRACT

This paper discusses the results of a recently completed flight test of an ITT VRS-1290 speaker dependent, continuous speech recognition system onboard a NASA Lewis Research Center OV-10A aircraft. A 54-word vocabulary was tested with thirteen pilots using an M-162 boom-mounted microphone on the ground and under 1g and 3g flight conditions. Digital audio tape (DAT) recordings were made of both the subjects' spoken phrases and the ambient background noise. Under some flight conditions, noise levels in the rear cockpit of the OV-10A were in excess of 115 dB which made it an ideal platform for testing the robustness of the ITT system. The DAT recordings were a critical element in optimizing the performance of the ITT system during the early stages of the flight test program. Average word accuracy for the thirteen pilots was 98.3% in the 1g condition and 97.3% in the 3g condition. Also discussed are plans for future testing using the DAT generated database with other speech systems.

## INTRODUCTION

The Pilot-Vehicle Interface Branch (FIGP) of Wright Laboratory (WL) is responsible for conducting basic, exploratory, and advanced research and development programs to design, implement and test advanced control and display technologies for integrated crew stations. Included in this research is the investigation of automatic speech recognition technology as a more natural and efficient means of controlling and managing various subsystems on-board the aircraft. Previous laboratory tests of this technology have shown a definite potential to provide significant speed and accuracy improvements over the use of knobs and switches in the cockpit (Warner and Harris, 1984; Enterkin, 1991). Flight test experiments conducted in the 80's provided the first opportunity to assess recognition performance of first generation airborne speech recognition systems (Werkowitz, 1984; Williamson and McDowell, 1986, Williamson, 1987). Results from these flight test programs suggested that improvements were needed before an operational system could be fielded in military aircraft.

Since that time significant progress has been made to advance the state-of-the-art in speech recognition technology. Continuous speech has replaced the more restrictive isolated speech systems tested almost a decade ago. It was time to reexamine the state-of-the-art to provide a new benchmark in robust performance with the current crop of available technology. The system chosen for this first round of evaluations was an ITT VRS-1290 Voice Recognizer Synthesizer system. This system had demonstrated excellent performance under quiet laboratory conditions (Barry, Solz, Reising, and Williamson, 1994).

In pursuing noise chamber testing with Armstrong Laboratory's Biological Acoustics branch (AL/CFBA), a unique flight test opportunity was discovered. AL/CFBA had been conducting flight experiments in active noise reduction and 3-D audio technologies on board a NASA Lewis Research Center (LeRC) OV-10A test aircraft. They suggested that WL might want to conduct a speech recognition evaluation on it as well. The OV-10 would be a good acid test for a speech system, with noise levels exceeding 120 dB in the rear cockpit. Discussions with NASA LeRC led to the signing of an Interagency Agreement in August 1994.

This paper discusses the flight testing of an ITT VRS-1290 speech recognition system in an OV-10A aircraft. In conducting the test, all speech was captured on high quality digital audio tape to be used for subsequent testing of other speech recognition systems.

## METHOD

### *Objective*

The objective of this experiment was to measure word recognition accuracy of the ITT VRS-1290 speech recognition system on the ground and in 1g and 3g flight conditions. A secondary objective was the collection of a speech database that could be used to test other speech recognition systems.

### Subjects

Sixteen subjects took part in this study. Twelve of the subjects were recruited from Wright-Patterson Air Force Base (WPAFB). All were rated military pilots. The remaining four subjects were NASA LeRC OV-10 pilots, each with prior military experience.

### Materials

The vocabulary consisted of 54 words that represent various tasks that could be accomplished in a military aircraft. The vocabulary and grammar structure are shown in Figure 1. The 54 vocabulary words were combined to form 91 test phrases to be used during ground and flight test conditions.

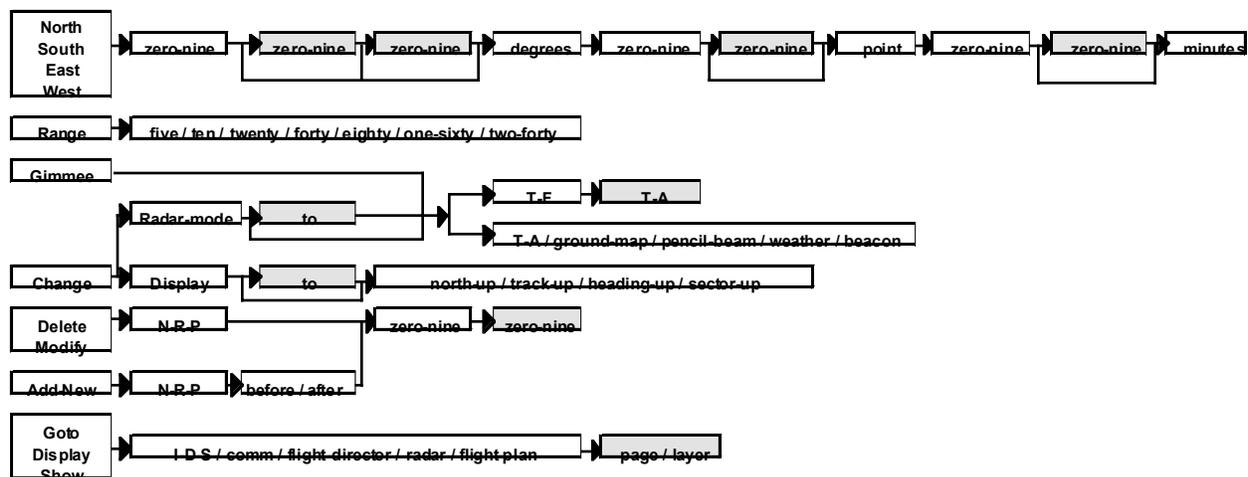


Figure 1. Vocabulary and syntax structure.

### Test Aircraft

The OV-10A aircraft used for this test was a twin engine, two crew member, tandem seating turboprop aircraft powered by two Garrett 1040 shaft horsepower engines. The aircraft had a minimum speed of 55 knots, maximum speed of 350 knots and a cruising speed of 185 knots. The aircraft had an endurance of approximately 2 hours.

### Flight Instrumentation & Test Hardware

A 90 inch long by 30 inch wide by 39 inch high cargo bay in the aircraft was used to house an avionics rack containing all research equipment. An IBM PC-compatible 80486 ISA bus computer system hosted the ITT VRS-1290 board and was mounted in the avionics rack along with a 24-track VHS data recorder. Figure 2 represents the instrumentation used during the flight test.

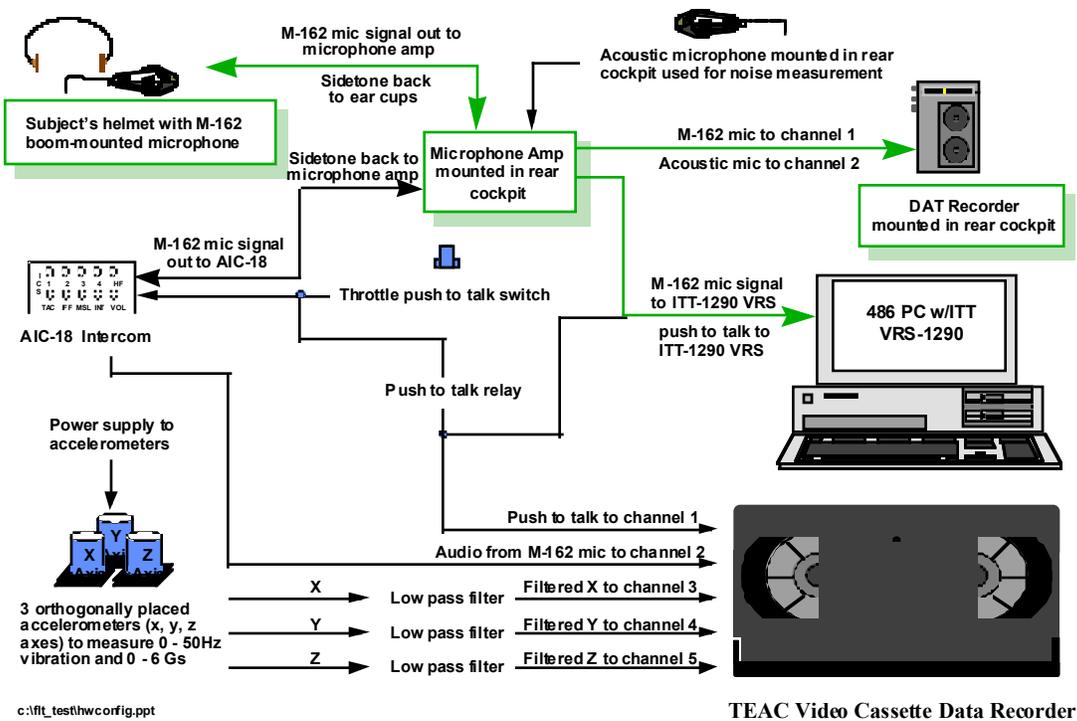


Figure 2. Flight Data Recording System

A DAT recorder was mounted in the rear cockpit to obtain high quality recordings of both the subjects' speech and the ambient noise inside the cockpit. These recordings will allow for future in-house testing of different speech recognition systems. Three (x, y, z) orthogonally arranged accelerometers, mounted on a block and placed close to the back seat of the aircraft, were used to measure both G and vibration experienced by the subject.

### Speech Recognition System

The ITT VRS-1290/PC Voice Recognizer Synthesizer system was used for all speech recognition tasks. The VRS-1290 is a speaker-dependent device which uses a Template Determined End Point (TDEP) speech processing algorithm to provide continuous speech recognition of up to 500 unique words at any one time, with a total capacity of 2,000 words. The system has an IBM PC AT/XT form factor with an operating range of 0-55 °C at 90% relative humidity. Although not intended for an airborne environment, the system functioned well in the rack-mountable PC located in the cargo bay of the OV-10A.

### Software

The ITT TGS (Template Generation System) program supplied with the speech recognition hardware was used to "train" the subjects (for template generation). Custom software written in-house was used for prompting the subjects, performing the speech recognition and subsequently recording the recognition results on the PC.

### Experimental Design

The experimental design was a single-factor Within Subjects design with five levels of the

Environment Independent Variable. All subjects were tested in the following conditions:

- 1) Lab - 182 phrases spoken in the laboratory environment
- 2) Hangar - 182 phrases spoken in the aircraft in the hangar with no engines running
- 3) 1g1 - 182 phrases spoken in the aircraft while flying wings level
- 4) 3g - 67 phrases spoken in the aircraft while pulling 3 gs
- 5) 1g2 - 182 phrases spoken in a second 1g condition to test for possible fatigue effects.

## PROCEDURE

### *Lab Testing*

Participation was divided into two separate sessions. The first session consisted of generating the subjects' templates in a laboratory setting and collecting some baseline performance data. Subjects were briefed on the nature of the experiment and performed template enrollment. An identical system to the one in the aircraft was used as the ground support system for template generation. The subjects used the same helmet and boom-mounted microphone that was used in the aircraft. Template training involved the subject speaking a number of sample phrases which were prompted by the TGS software package delivered with the ITT hardware system. Once template generation was completed, a recognition test followed which consisted of reciting the phrases to collect baseline recognition data. Each of the 91 vocabulary phrases were spoken twice for a total of 182 phrases spoken in the laboratory. All of the laboratory training and testing utterances were recorded to allow subsequent testing on the ITT system or testing of a new speech recognition system.

### *Aircraft Testing*

The second session began with a cockpit briefing that covered the operation of the test equipment (starting the DAT recorder, placement of the microphone, etc.), and safety issues. The subjects were provided with a knee board that contained the various checklists and a printout of the phrases to be spoken during the flight test. This printout was provided as a back-up in case of equipment problems.

During data collection, both on the ground and in the air, subjects sat in the rear seat of the OV-10A and were prompted with a number of phrases to speak. All prompts appeared on a 5" x 7" monochromatic liquid crystal display in the instrument panel directly in front of the subject. The recognition system attempted recognition after each spoken phrase with the recognition result stored for later analysis. DAT recordings were made of the entire data collection session.

The first aircraft test session was performed in the hangar to provide a baseline on the aircraft in quiet conditions. This consisted of each subject speaking the 91 test phrases twice -- for a total of 182 phrases. During both ground and airborne testing, subjects needed little or no assistance from the pilot of the aircraft. Close coordination was required, however, between the pilot and subject while the 3g maneuvers were being performed.

The flight test profile consisted of three conditions: (1) straight and level flight (1g), (2) 3g flight, and (3) repetition of the 1g condition to examine potential fatigue effects.

*Straight and level flight (1g):* After leveling-off (5000 to 10000 feet) the pilot maintained a constant power setting (approximately 85%) with the aircraft in the clean configuration (landing gear and flaps retracted). The subjects spoke 182 phrases in the 1g condition initially (two repetitions of all 91 phrases), and 182 phrases in the 1g condition after the 3g maneuvers. This resulted in a total of 364 phrases in the 1g condition.

*3g maneuvers:* After completing 182 phrases in the 1g condition, the pilot began a series of 3g

maneuvers (720° turns, 70° bank, power setting of 85%). The subjects were required to speak 67 phrases during these maneuvers.

### *Performance Optimization*

During the course of data collection, several problems arose which resulted initially in poor recognition performance. These problems were primarily audio related but also had to do with a lack of understanding of some of the engineering parameters that controlled the ITT system. Two such parameters were Noise Tracker and Noise Tracker Rejection flags that were both enabled. These parameters were primarily designed to enable rejection of spurious impulse noises, such as door slams. With the noise tracker parameters enabled, the system too often rejected valid utterances as noise, especially utterances at a terminal node of the grammar. Disabling these parameters resulted in at least a 10% improvement in word accuracy.

Another problem occurred when the subjects were required to perform a calibration of the system prior to a given flight condition. This process performed two functions simultaneously: background noise calibration and automatic gain control (AGC) parameter setting. During noise calibration, the system prompted each subject to be quiet for a short period. During this silence period, the system would generate a template of the background noise to adjust the voice templates for use in the higher noise environment. Due to procedural problems, however, this noise calibration was sometimes bypassed, resulting in poor inflight recognition performance. During AGC adjustment, the subjects were required to speak the phrase “ONE TWO THREE FOUR”. After the digit phrase was spoken, the system would adjust the gain up or down and repeat the process until it was satisfied with the audio level. Most of the time, however, the system would freeze during this calibration step, requiring the subject to restart the computer. In order to ensure an accurate noise reading, the gain had to be fixed at a certain level. This required extensive retest with the DAT tapes generated in flight to find an optimum gain setting. It was discovered during the course of gain optimization that the flight system was being overdriven. Once the input audio gain to the ITT was reduced, performance again improved dramatically.

## **RESULTS**

Due to the audio and system problems encountered during the experiment, only five of the sixteen subjects had valid real-time recognition performance data in-flight. Three of the sixteen subjects experienced problems with the DAT recording equipment, resulting in unusable or non-existent audio data. Audio recordings were successfully collected for a total of thirteen subjects in the study.

The data analyses were done in two stages. The first stage involved a comparison of “live”, in-flight word recognition performance with word recognition performance obtained by playing the DAT recordings made in-flight into the ITT system back in the laboratory. The premise was that if no significant differences were found between live vs. DAT performance on the five subjects that flew with the optimum configuration, then the remaining subjects with complete DAT audio could be retested in the lab in the same way. Table 1 shows the mean word recognition performance for both live and DAT recordings for the five subjects who had valid in-flight data.

Condition	1g1	3g	1g2
Live	98.18%	98.25%	98.17%
DAT	98.31%	98.28%	98.48%

Table 1. Mean word accuracy for live and DAT testing.

An Analysis of Variance revealed no significant differences in word recognition performance when providing the ITT system with both live and digitally recorded audio signals.

With no performance differences found between live and DAT audio signals, all of the remaining analyses were done using DAT audio tape as the input to the VRS-1290. This provided complete recognition data for thirteen subjects. Table 2 shows the mean word recognition performance obtained for each of the test conditions.

Recognition Performance	Lab	Hangar	1g1	3g	1g2
	98.24%	98.42%	98.55%	97.3%	98.15%

Table 2. Mean word accuracy for each test condition.

Three comparisons were of primary interest:

1. Ground (Lab + Hangar) versus air (1g1 + 3g + 1g2) performance
2. 1g (1g1 + 1g2) versus 3g performance
3. 1g1 versus 1g2 performance

Orthogonal comparisons were done to make each of these comparisons. No significant differences were found for any comparisons.

## DISCUSSION

Once the audio level and system parameters were optimized, the ITT VRS-1290 Voice Recognizer Synthesizer system performed very well, achieving over 98% accuracy over all flight conditions. It was anticipated that performance would more significantly degrade in flight, especially in the 3g condition, due primarily to the increase in noise level during this maneuver. This, however, was not the case. Once the background noise calibration was performed, the system was able to effectively compensate for the aircraft noise background.

This experiment highlighted several important lessons learned in the flight testing of automatic speech recognition systems. First, audio is everything. If the AGC function on the ITT VRS-1290 performed as it should, the system would have adjusted the input gain to the appropriate level. Second, the speech application designer needs to be very familiar with the various parameters that control portions of the recognition process. In this case, changing a Noise Tracker flag from a 1 to a 0 reduced word error rate by over 10%.

## CONCLUSION

This flight test represented one of the most extensive in-flight evaluations of a speech recognition system ever performed. Over 5600 phrases comprised of over 27,000 utterances were spoken by the thirteen subjects in flight. This combined with the two ground conditions resulted in a test of over 51,000 utterances. The audio database of DAT recordings will be transferred onto CD-ROM to facilitate laboratory testing of other speech recognition systems. Of particular interest is to determine if any of the currently available software-based speaker-independent systems are capable of achieving the same level of performance as the speaker-dependent ITT system. The CD-ROM database will also be available for distribution to the speech recognition research community.

## REFERENCES

Barry, T. P., Solz, T. J., Reising, J. M., and Williamson, D. T. (1994). The use of word, phrase, and intent accuracy as measures of connected speech recognition performance. In *Proceedings of the*

*Human Factors Society 38<sup>th</sup> Annual Meeting* (pp. 325-329). Nashville, TN: Human Factors Society.

Enterkin, P. (1991). Voice versus manual techniques for airborne data entry correction. In *Proceedings of the Ergonomics Society 1991 Annual Conference, Ergonomics - Design for performance 1991*. London: Taylor & Francis.

Warner, N. and Harris, S. (1984). Voice-controlled avionics programs, progress and prognosis. In *Proceedings of Speech Tech '84 Voice Input/Output Applications Show and Conference* (pp. 110-123). New York, New York: Media Dimensions.

Werkowitz, E. B. (1984). Speech recognition in the tactical environment: the AFTI/F-16 voice command flight test. In *Proceedings of Speech Tech '84 Voice Input/Output Applications Show and Conference* (pp. 103-105). New York, New York: Media Dimensions.

Williamson, D. T. and McDowell, J. W. (1986). The implementation of a voice actuated radio management system in a C-135 aircraft. In *Proceedings of AVIOS '86 Voice I/O Systems Applications Conference* (pp. 144-151). Alexandria, VA: American Voice Input/Output Society.

Williamson, D. T. (1987). Flight test results of the AFTI/F-16 voice interactive avionics program. In *Proceedings of AVIOS '87 Voice I/O Systems Applications Conference* (pp. 335-345), Alexandria, VA: American Voice Input/Output Society.