

**Robust Speech Recognition Interface to the Electronic Crewmember:
Progress and Challenges**

David T. Williamson

**Vehicle-Pilot Integration Branch
2210 Eighth St Suite 11
Wright-Patterson AFB, OH 45433-7521**

SUMMARY

Speech is a natural form of communication between humans. It should come as no surprise that it would also be the ideal form of communication between a pilot and an electronic crewmember. High-level commands spoken by the pilot would be interpreted and carried out by the electronic crewmember in much the same way that a pilot would talk to another crewmember. The realization of this natural interface will depend on a robust speech recognition capability to handle the degraded speech conditions typical of the military aircraft environment. This paper reviews the latest progress in robust speech recognition research and its potential application for military aircraft. Sources of degradation in the speech signal will be discussed along with the techniques being explored to reduce their effects on speech recognition. Results of recent flight testing will also be presented to provide a benchmark of the performance of commercially available speech systems in the military environment. Finally, remaining challenges to providing a fully capable, high-accuracy speech interface to the electronic crewmember will be discussed.

1 INTRODUCTION

Speech technology holds the promise of providing a natural means for human crewmembers to communicate with their future electronic counterparts. Automatic speech recognition is a rapidly emerging human-computer interface technology that will provide a safe and efficient method for handling the complex information management requirements of future fighter aircraft. The Vehicle-Pilot Integration Branch in Wright Laboratory (WL) has been actively investigating the potential of this technology for over twenty years. Flight test experiments conducted in the 80's provided the first opportunity for WL to assess recognition performance of first generation airborne speech recognition systems (ref. 1, 2). Results from these early flight test programs suggested that significant improvements in the area of robust speech recognition were needed before an operational system could be fielded in military aircraft.

Robustness refers to the ability of a speech recognition system to operate under adverse conditions. In the military environment these adverse conditions are concentrated in two areas: noise and speech variability. Noise is produced by the aircraft engines, wind, environmental control system, oxygen mask breath noise, and electrical channel noise produced by distortion in the

microphone and avionics systems. Speech variability is primarily caused by g forces, workload stress, fatigue, and Lombard speech. Lombard speech occurs when speakers attempt to make themselves heard over the background noise. If speech recognition is to be a viable cockpit information management technology, researchers must fully understand the factors that degrade the speech signal in the operational environment and develop robust algorithms to compensate for them.

Fortunately, significant progress has been made in robust speech recognition since those first flight test experiments in the 80's. Improvements in digital signal processing technology are resulting in relatively inexpensive speech recognition systems that are designed to operate in noisy industrial environments for command-and-control and data entry applications. Also, the growing market for computer telephony applications is resulting in technology capable of recognizing speech over noisy telephone lines. With little modification, a system designed for these commercial applications could be adapted for use in military aircraft.

This paper reviews the latest progress toward achieving a robust speech recognition interface for military aircraft. Sources of degradation in the speech signal will be discussed along with the techniques being explored to reduce their effects on speech recognition. Results of two recent WL flight tests on two NASA OV-10 aircraft will also be summarized to provide a performance benchmark of commercially available speech systems in the airborne environment. Finally, remaining challenges to providing a fully capable, high-accuracy speech interface to an electronic crewmember will be discussed.

**2 SOURCES OF COCKPIT SPEECH
DEGRADATION**

As mentioned above, there exists two primary areas that can degrade the speech signal in the military aircraft environment: noise and speech variability. Each of these sources along with ways of compensating for them is discussed below.

Noise Sources

Three major noise sources that contribute to degradation of the speech signal in a cockpit are ambient background noise, channel noise, and speaker noise. Ambient background noise is produced by the aircraft engines, environmental control system, and the sound of air moving past the aircraft. Channel noise refers to the

distortions produced by the microphone transducer and electrical noise conducted in the wiring to the speech system. Speaker noise refers to non-vocabulary speech sounds such as lip smacks, breath noise from an oxygen mask, or grunting sounds produced when a pilot undergoes high-g maneuvers.

Three of the most commonly used techniques for reducing the effects of these noise sources are training in the environment, preprocessing, and noise cancellation algorithms. If the noise source is relatively consistent and steady-state, as it is for many airborne environments, then having the speaker train the system in noise conditions that simulate the actual aircraft environment is one method of producing representative voice templates that will more closely match commands given in flight. Preprocessing and noise cancellation are preferred over training in noise, however, as these techniques try to reduce noise effects by modification of the speech signal rather than requiring additional training from speakers (ref. 3, 4).

Oxygen mask breath noise is a unique problem that has to be dealt with for high-performance aircraft applications. The creation of breath noise models was successfully used with the ITT VRS-1290 speech system during the second OV-10 flight test experiment conducted by Wright Laboratory and will be discussed in the next section. Training the system with the oxygen mask also incorporates the intra-utterance breath noise as part of the word models and minimizes its impact.

Other speech noises such as lip smacks, grunts, or out-of-vocabulary utterances are more difficult to deal with. One method is to adjust the recognition threshold of a system to reject out-of-vocabulary sounds. The problem with doing this, however, is that there is a danger of the system rejecting too much valid speech as well. Another method is to incorporate "garbage" models that recognize and reject speech noises and out-of-vocabulary speech. This is typical of word-spotting systems that can recognize keywords in a continuous stream of speech.

Speech Variability

The second major area of speech degradation occurs when the pilot's voice changes due to various factors such as g forces, workload stress, fatigue and the Lombard effect. In previous flight test experiments, flying up to 6 g's resulted in little degradation in speech recognition performance (ref. 2). Fortunately, there is very little application for speech recognition above 6 g's and with the proper closed-loop feedback of g level to the speech system, the vocabulary and grammar structure can be significantly limited to only enable those few tasks that a pilot may want to access by voice.

The effect of background noise alone on speech recognition performance is not as detrimental as how the noise affects a speaker's response to it. This Lombard effect, named after the French physician who first described its characteristics (ref 5), results in changes to a speaker's voice such as increased vocal effort, greater duration of words due to an elongation of vowels,

frequency shifts, and deletion of certain ending consonants. While this effect makes it easier for humans to communicate in noise, it can reduce speech recognition accuracy by as much as 25 percent. The best techniques for minimizing Lombard speech are providing good audio feedback in the speaker's headset, minimizing the noise through the use of active noise reduction, and feedback techniques that provide the speaker with the gain level the system is receiving (ref. 6, 7, 8).

3 OV-10 SPEECH RECOGNITION FLIGHT TESTING

To assess the impact of these various sources of speech degradation on commercially available speech recognition systems, two flight test experiments were recently conducted by WL on two NASA Lewis Research Center OV-10 test aircraft. The objectives of these experiments were 1) measure live recognition performance in several ground and flight test conditions, including testing up to 4g's and 2) generate a digital speech database for further research.

Experiment 1 - ITT VRS-1290 Evaluation with an M-162 Boom Microphone

Sixteen subjects, comprised of active duty military and NASA pilots, participated in the evaluation of an ITT VRS-1290 speech recognition system installed in a ruggedized IBM-PC. The aircraft used for this experiment was an OV-10A aircraft operated by NASA Lewis Research Center in Cleveland, OH (Figure 1). This aircraft was a twin engine, two crew member, tandem seating turboprop aircraft. The OV-10A was capable of pulling up to 5.5 gs, but due to equipment constraints the test profiles were limited to 1 and 3g maneuvers.



Figure 1. NASA LeRC OV-10A Test Aircraft

Vocabulary/Grammar Structure

The vocabulary consisted of 53 words and phrases that represent various tasks that could be accomplished in a military aircraft. The vocabulary and grammar structure is shown in Table 1. The 53 vocabulary words and phrases were combined to form 91 test utterances to be used during ground and flight test conditions. Synonymous words such as Go-to, Display, and Show or page and

layer were designed into the test vocabulary to allow a more flexible interaction with the speech system.

- {North/South/East/West} (0 - 1 8 0) degrees (0 - 5 9) point (0 - 9 9) minutes**
- Range {five/ten/twenty/forty/eight/one-sixty/two-forty}**
- Give-me {TF/TA/TF-TA/ground-map/pencil-beam/weather/beacon}**
- Change Radar-mode [to] {TF/TA/TF-TA/ground-map/pencil-beam/weather/beacon}**
- {Delete/Modify} N-R-P (0 - 9 9)**
- Add-New N-R-P {before/after} (0 - 9 9)**
- {Goto/Display/Show} {IDS/comm/flight-director/radar/flight-plan} {page/layer}**

Table 1. Flight Test Vocabulary/Grammar

Test Procedures

Each subject began the experiment by performing template generation followed by a baseline performance assessment. Template generation involved the subjects' speaking a number of sample utterances which were prompted by the ITT system. Once template generation was completed, a recognition test followed which consisted of reciting 91 utterances twice to collect baseline recognition data. All of the laboratory training and testing utterances were recorded on digital audio tape (DAT) to allow subsequent testing on the ITT system or testing of a new speech recognition system.

The subsequent test sessions were conducted on the aircraft both on the ground with no engines running and in the air. During data collection, subjects sat in the rear seat of the OV-10A and were prompted with a number of utterances 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 ITT system attempted recognition after each spoken phrase with the results stored for later analysis. Once again, DAT recordings were made of the entire data collection session. After the ground test was complete, the subjects flew the flight test profile consisting of three conditions: 1) straight and level flight (1G1), 2) 3g flight (3G), and 3) repetition of the 1g condition to examine potential fatigue effects (1G2).

Results

During the first several flights, ITT word accuracy was around 55%. In the course of investigating potential causes for this performance degradation, several problems were discovered. These problems were primarily audio related but also had to do with several engineering parameters that controlled the ITT system. After consultation with ITT researchers, DAT flight test recordings were replayed into the system on the ground with systematic adjustments to the gain and engineering parameters. Recognition performance was then obtained at greater than 98%. Once this performance optimization

was accomplished, live performance was maintained at 98% or better across all flight conditions with no significant degradation at 3g's.

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. Four 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 twelve 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. Figure 2 shows the mean word recognition performance for both live and DAT recordings for the five subjects who had valid in-flight data.

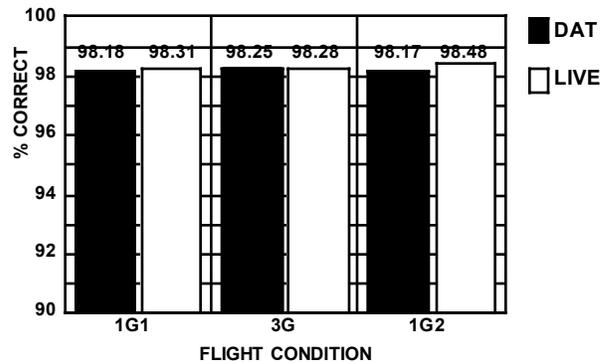


Figure 2. Mean word accuracy for live and DAT testing

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 twelve subjects. Figure 3 shows the mean word recognition performance obtained for each of the test conditions. Statistical analysis revealed no significant differences in any of the test conditions.

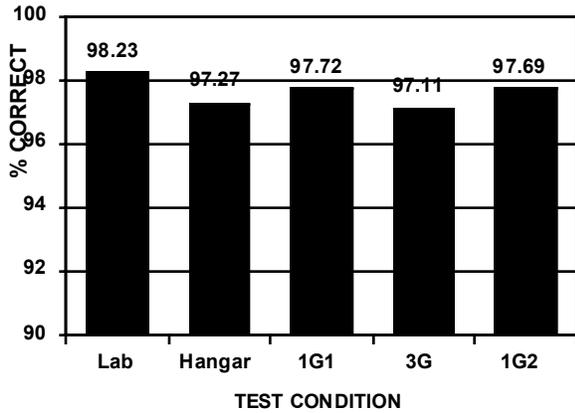


Figure 3. Mean word accuracy for each test condition.

Experiment 2 - ITT VRS-1290 and Verbex VAT31 Evaluation with an M-169 Oxygen Mask Microphone

A second experiment was conducted, this time using an oxygen mask with an Air Force standard M-169 microphone to examine the effects of aircraft noise, breath noise and g's on speech recognition performance. Ten subjects participated in the first stage of the experiment which evaluated the ITT VRS-1290 speech recognition system, this time installed on a NASA OV-10D aircraft. After the ITT testing was completed, a Verbex VAT31 was installed and evaluated with six subjects using the same vocabulary and grammar structure. Since different subjects were used for both systems, a direct comparison between the ITT and Verbex was not performed. Also, both ITT and Verbex were consulted to ensure optimum performance for both systems.

Vocabulary/Grammar Structure

A new vocabulary and grammar structure was developed for this experiment. The vocabulary consisted of 47 words and phrases, some of which were used during the second AFTI/F-16 flight test that was performed over ten years ago (ref. 2). The vocabulary and grammar structure is shown in Table 2. A total of 57 test utterances was developed to be used during ground and flight test conditions.

- (Uniform/Comm 1) (2 0 0 0 - 3 9 9 9)
- (Victor/Comm 2) (1 5 0 0 - 1 9 9 9)
- (Uniform/Comm 1) (button/channel) (1 - 2 0)
- (Victor/Comm 2) (button/channel) (1 - 2 0)
- Radar range (ten/twenty/forty/eighty)
- Radar azimuth (ten/thirty/sixty)
- Radar (1/2/3/4) bar
- (HI-TACAN/ILS) runway (0 0 - 3 6) (Left/Right)
- (Say/Whats-my) (fuel/inventory)
- (Tracker/Targeting-pod) (wide/narrow/cursor-zero)
- (Tracker/Targeting-pod) (black/white) hot
- CCIP
- Air-to-air-mode
- Air-to-Ground-mode
- Nav-mode
- Strafe

Table 2. OV-10D Flight Test Vocabulary

Test Procedures

The test procedures for both the ITT and Verbex systems were repeated from the first experiment. Each subject began the experiment by performing template generation followed by a baseline performance assessment. After that, a recognition test followed which consisted of reciting 57 utterances twice to collect baseline recognition data. Once again, all of the training and testing utterances were recorded on DAT to allow follow-on testing.

The subsequent test sessions were conducted on the aircraft both on the ground with no engines running and in the air. After the 114 utterance ground test was complete, the subjects flew the flight test profile consisting of three conditions: 1) 114 utterances at straight and level flight (1G1), 2) 70 utterances in 4g flight (4G), and 3) 114 utterances repeating the 1g condition at a higher engine throttle setting to induce more noise for this condition. (1G2).

Results - ITT Testing

Due to various data recording and aircraft problems, DAT audio was successfully recorded for only eight of the ten subjects under all test conditions. Live recognition performance was obtained for five of these eight subjects. Figure 4 shows the word accuracy for five subjects under each test condition. Two factors accounted for the majority of the recognition errors, lack of automatic gain control and Lombard effect. As a result of the first flight test, the ITT system was used without its automatic gain control circuitry enabled. This was because the ITT system had difficulty converging on the proper gain setting when exposed to high noise. Also, when it finally did settle on a particular gain setting, it was found that the signal was too strong to obtain accurate results. So fixing the gain to a predetermined value was the only way to get the ITT system to function reasonably well in this second flight test.

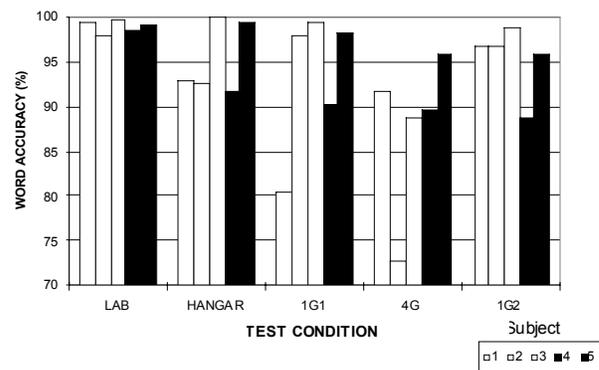


Figure 4. ITT word accuracy for five subjects

This became a problem, however, with subjects that had a pronounced Lombard effect, particularly in the 4G condition. During the 4G run, the noise level increased by an average of 22 dB from the 1G1 condition. This, coupled with a lack of good sidetone in the subjects' helmet earcups, resulted in some subjects almost shouting to compensate for the increased noise level. This explains

why subject 2's word recognition results at 4G were at 72.7%. Subject 5, however, was very accustomed to speaking in the OV-10 and consequently was able to maintain performance at 95% or better for all conditions. Average performance over the five subjects was 97.2% for the two ground conditions and 92.1% for the three flight conditions. Subsequent experiments are planned with the DAT audio to determine if gain normalization will improve ITT performance.

Results - Verbex Testing

Five of the six subjects had complete live data and DAT audio for each of the five test conditions. Figure 5 shows the live word recognition performance for each of the five subjects. Subject 5 is the same subject as subject 5 in the ITT test. Due to his experience with both the aircraft and the testing procedures, his performance was the best at 100% under all conditions. Subject 3 was a non-pilot subject that showed a pronounced Lombard effect under 4g's. This explains the performance degradation of 86% in the 4G condition. Overall, the system achieved an average word accuracy of 99.5% in the ground conditions and 97.3% in the flight conditions.

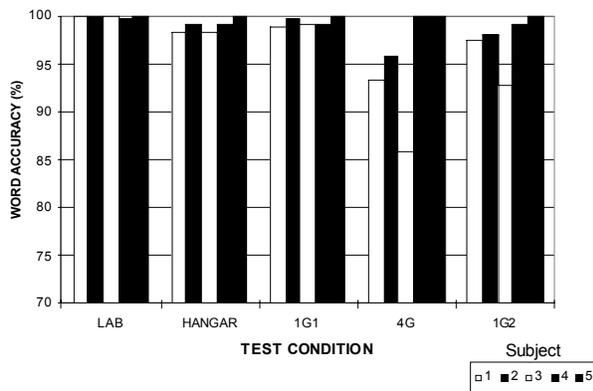


Figure 5. Verbex word accuracy for five subjects

OV-10 Flight Test Conclusions

The two flight test programs summarized here provided an excellent opportunity to obtain practical experience with the airborne evaluation of commercially available speech recognition systems. Perhaps of greater significance than the actual recognition results, however, is the fact that an extensive digital speech database was recorded that will be distributed to other speech recognition researchers to develop and evaluate recognition algorithms for the airborne environment. This database will also be used internally to evaluate other candidate speech systems without going through the expense of additional flight testing. Of particular interest is the evaluation of several speaker independent systems that do not require training prior to use.

4 FUTURE DIRECTIONS AND CHALLENGES

A robust speech interface in the cockpit is fast becoming a cost effective technology option for crew systems designers. With digital signal processing speed

increasing about 20 percent each year, the necessary horsepower required to provide high accuracy, real-time speech recognition in the military environment is already here for small vocabulary, continuous speech command and control applications. With the latest push to adopt commercial technology for military use, the Air Force can leverage a tremendous investment by commercial developers working on robust speech interfaces to automobile systems, cellular telephone dialing, information kiosks, personal digital assistants, etc. While none of these environments can fully compare with the operational fighter environment, technology gains made in the private sector will have direct application to the military.

Several new approaches to improving robust speech recognition are also showing a lot of promise in the laboratory. Researchers are exploiting the use of neural networks for improved pattern recognition and auditory modeling techniques that mimic the excellent noise filtering characteristics of the human auditory system (ref 9, 10).

Once this robust speech processing capability is available, the remaining challenges lie in the application designer developing a natural, intuitive interface between the pilot and the electronic crewmember. Speech understanding systems that are able to interpret meaning from spontaneous conversational speech input are still in their infancy. Fortunately, fighter pilots have little need for verbose discourse with their aircraft and would rather communicate in very short, unambiguous commands. No other human-computer interface technology has the potential for providing as rapid and efficient an interface, allowing the pilot to respond to mission events at a higher level of control and provide the timely decision support needed to return home safely.

5 REFERENCES

1. 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.
2. 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.
3. Hermansky, H., Morgan, N., and Hirsh, H. (1993). Recognition of speech in additive and convolutional noise based on RASTA processing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 2.* (pp. 83-86).
4. Hansen, J.H.L. (1993). Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 2.* (pp. 95-98).

5. Lombard, E. (1911). Le signe de l'elevation de la voix. In *Ann. Maladies Orielle, Larynx, Nez, Pharynx*. 37:101-119.
6. Fletcher, H., Raff, G., and Parmley, F. (1918). Study of the effects of different amounts of sidetone in the telephone set. Technical Report 19412, Western Electric Company.
7. Lane, H., Tranel, B., and Sisson, C. (1970). Regulation of voice communication by sensory dynamics. In *Journal of Acoustical Society of America*, 47(2):618-624.
8. Pick, H., Siegel, J., Fox, P., Garber, S., and Kearney, J. (1989). Inhibiting the Lombard effect. In *Journal of Acoustical Society of America*, 85(2):894-900.
9. Cheng, Y. M. and O'Shaughnessy, D. (1991). Speech enhancement based conceptually on auditory evidence. In *Transactions on Speech Processing*, 39(9): 1943-1946.
10. DeSimio, M. and Anderson, T. (1993). Phoneme recognition with binaural cochlear models and the stereausis representation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing I*. (pp. 521-524).