DRDC Toronto CR-2005-033

VOICE CONTROLS AND DISPLAYS FOR THE DISMOUNTED SOLDIER

by:
L. Thompson and D. Tack

Human*systems*® Incorporated
111 Farquhar St., 2nd floor
Guelph, ON   N1H 3N4

Project Manager:
David W. Tack
(519) 836 5911

PWGSC Contract No. W7711-017747/001/TOR
Call-Up 7747-01
HSI® SIREQ Item #90

On behalf of
DEPARTMENT OF NATIONAL DEFENCE

as represented by
Defence Research and Development Canada - Toronto
1133 Sheppard Avenue West
Toronto, Ontario, Canada
M3M 3B9

DRDC Toronto Scientific Authority
Maj Linda Bossi
(416) 635-2197

October 2005

This document contains information that may not be passed or shared, even in confidence, with foreign military, research and development representatives or civilian contractors of any nationality without the expressed prior permission of the Exploitation Manager of SIREQ TD.

The scientific or technical validity of this Contract Report is entirely the responsibility of the contractor and the contents do not necessarily have the approval or endorsement of Defence R&D Canada

© Her Majesty the Queen as represented by the Minister of National Defence, 2005

© Sa Majesté la Reine, représentée par le ministre de la Défense nationale, 2005

Abstract

The purpose of this report is to investigate the potential application of voice recognition and speech synthesis to a wearable computer for use by dismounted soldiers. A literature review was conducted to identify the issues and capabilities of voice controls and displays for dismounted infantry soldiers. Current applications of voice interfaces are identified including the Force XXI modified Land Warrior user voice control system. Considerations regarding the appropriate use of speech input and voice displays are discussed including use of speech as opposed to tonal or visual displays, text-to-speech synthesis, types of speech displays and speech display intelligibility. Factors influencing the effectiveness of speech input and voice control systems are described including speaker dependence, word boundaries, vocabulary size, and system performance. Environmental factors affecting speech recognition accuracy are discussed including ambient noise, vibration, and operator stress. Finally, this report considers general system design issues including speech input device activation and deactivation, multi-modal displays and controls, feedback and error correction, vocabulary size and selection, robustness to ambient noise, system size, microphone and headset design, compatibility with masks and exclusion garments, training, and security issues.

Résumé

L'objet du présent rapport est d'étudier l'application éventuelle de la reconnaissance vocale et de la synthèse de la parole à un ordinateur vestimentaire destiné à être utilisé par les fantassins débarqués. On a réalisé une analyse documentaire afin de relever les problèmes que posent les systèmes de commande et d'affichage vocaux aux fantassins débarqués ainsi que les capacités qu'ils leur offrent. On fait état des applications actuelles des interfaces vocales, y compris du système à commande vocale du fantassin du 21^{e} siècle. On évoque les aspects concernant l'utilisation appropriée de l'entrée vocale et des systèmes d'affichage vocaux et notamment l'usage de la parole par opposition aux tonalités et aux affichages visuels, la synthèse de la parole à partir du texte, les types de systèmes d'affichage vocaux ainsi que l'intelligibilité de ces derniers. On y décrit les facteurs qui ont une incidence sur l'efficacité des systèmes à entrée vocale et à commande vocale, y compris la dépendance du locuteur, les frontières de mots, la richesse lexicale ainsi que les performances du système. On discute également des facteurs propres au milieu qui ont des répercussions sur le niveau d'exactitude en matière de reconnaissance vocale, y compris les bruits ambiants, les vibrations et le stress de l'opérateur. Le présent rapport s'intéresse enfin à des problèmes de conception de systèmes généraux au rang desquels figurent l'activation et la désactivation des systèmes à entrée vocale, les systèmes de commande et d'affichage combinés, la rétroaction et la correction des erreurs, la richesse lexicale et la sélection du vocabulaire, la résistance aux bruits ambiants, la taille du système, la conception des microphones et des casques, la compatibilité avec les masques et les vêtements filtrants, la formation et les problèmes en matière de sécurité.

Executive Summary

Continuous technology improvements over the years have now allowed speech to be considered as a feasible input and output medium between soldiers and computers. Various studies have been performed to investigate the application of voice displays and controls to many different domains, but its application to future dismounted infantry operations is not well understood. The suitability of employing a voice interface in dismounted infantry operations, and the associated issues, will be evaluated as part of the Solder Information REQuirements Technology Demonstration (SIREQ TD) project.

This technical report investigates the potential application of voice recognition and speech synthesis to a wearable computer for a dismounted soldier to facilitate hands-busy, eyes-busy control of equipment and personnel by voice. Through an extensive literature review, the issues and capabilities of voice controls and displays are discussed. Current applications of voice interfaces are identified including the Force XXI modified Land Warrior user voice control system. Considerations regarding the appropriate use of speech output and voice displays are discussed including use of speech as opposed to tonal or visual displays, text-to-speech synthesis, types of speech displays, and speech display intelligibility. Factors influencing the effectiveness of speech input and voice control systems are described including speaker dependence, word boundaries, vocabulary size, and system performance. Environmental factors affecting automatic speech recognition, along with issues in the design of a voice control and display system for the soldier are also discussed. Current and future military applications of voice interfaces are noted along with a discussion of the current commercial technology available to use for laboratory and field testing.

This report concludes that although voice controls and displays can be an effective information control and display interface for the dismounted soldier, their limitations and capabilities must be taken into consideration. Voice displays are best used when the soldier is moving and the operational conditions do not allow the use of a visual display. Speech presentation is best for short, simple messages that do not need to be referred to later in time. Voice displays require a text-to-speech synthesizer to produce the speech messages through phonemes or recorded text. Users prefer when the synthesized voice is distinguishable from other human voices. Current displays are usually menu or keyword driven, but as the technology improves, a natural language interface is possible. Altering tones or cue words, preceding the voice message, can be used to add additional information to a speech message. Speech displays are well suited to warnings, prompts, feedback, and short responses to user queries. General design considerations for the voice input/output system for the soldier's computer are also discussed.

Sommaire

Les améliorations technologiques continuelles apportées au fil des ans ont fait en sorte que la parole constitue désormais un moyen d’entrée et de sortie envisageable entre les soldats et les ordinateurs. Si plusieurs études ont été menées sur l’application des systèmes de commande et d’affichage vocaux à de nombreux domaines différents, leur application aux opérations de l’infanterie débarquée, dans l’avenir, n’est pas bien comprise. Dans le cadre du Projet de démonstration technologique des besoins des soldats en matière d'information (SIREQ TD), on évaluera dans quelle mesure il serait possible d’avoir recours à un système d’interface vocal pour les fins des opérations de l’infanterie débarquée ainsi que les problèmes connexes que cela soulève.

Le présent rapport technique s’intéresse à l’application éventuelle de la reconnaissance vocale et de la synthèse de la parole à un ordinateur vestimentaire destiné aux fantassins débarqués, l’objectif étant de faciliter l’utilisation de l’équipement et la commande du personnel, par la voix, alors que tant les mains que les yeux sont déjà occupés à d’autres tâches. On évoque les problèmes propres aux systèmes de commande et d’affichage vocaux ainsi que leurs capacités, par le biais d’une analyse documentaire approfondie. On fait état des applications actuelles des interfaces vocales, y compris du système à commande vocale du fantassin du 21^{e} siècle. Il est question de l’utilisation appropriée de la sortie vocale et des systèmes d’affichage vocaux et notamment de l’usage de la parole par opposition aux tonalités et aux affichages visuels, de la synthèse de la parole à partir du texte, des types de systèmes d’affichage vocaux ainsi que de l’intelligibilité de ces derniers. On y décrit les facteurs qui ont une incidence sur l’efficacité des systèmes à entrée vocale et à commande vocale, y compris la dépendance du locuteur, les frontières de mots, la richesse lexicale ainsi que les performances du système. On discute également des facteurs propres au milieu qui ont des répercussions sur la reconnaissance vocale automatique ainsi que des problèmes liés à la conception des systèmes de commande et d’affichage vocaux destinés aux soldats. On fait état des applications militaires actuelles et futures des interfaces vocales, en plus de discuter des technologies commerciales actuellement disponibles à des fins d’essai en laboratoire et sur le terrain.

Dans le rapport, on en vient à la conclusion que, bien que les systèmes de commande et d’affichage vocaux puissent constituer une interface d’affichage et de commande de l’information efficace pour les fantassins débarqués, il convient de tenir compte de leurs limites et de leurs capacités. Les systèmes d’affichage vocaux s’avèrent particulièrement utiles lorsque les soldats se déplacent et que les conditions opérationnelles ne permettent pas d’avoir recours à un système d’affichage visuel. Les systèmes de synthèse de la parole conviennent particulièrement aux messages courts et simples auxquels il n’est pas nécessaire de se référer ultérieurement. Les systèmes d’affichage vocaux nécessitent un synthétiseur de la parole à partir du texte pour produire les messages vocaux en s’appuyant sur des phonèmes ou du texte enregistré. Les utilisateurs préfèrent les systèmes qui permettent de distinguer la voix synthétisée des autres voix humaines. Les systèmes d’affichage actuels sont généralement pilotés par menus ou par mots clés, mais les progrès technologiques permettent d’entrevoir le recours à une interface en langage naturel. On peut faire précéder le message vocal de tonalités différentes ou de mots clés pour ajouter des renseignements complémentaires à un message vocal. Les systèmes d’affichage vocaux conviennent bien aux messages d’avertissement, aux invites, aux commentaires ainsi qu’aux réponses courtes aux interrogations de l’utilisateur. On traite aussi des considérations générales en matière de conception des systèmes vocaux d’entrée et de sortie destinés aux ordinateurs des soldats.

Table of Contents

ABSTRACT I

RÉSUMÉ II

EXECUTIVE SUMMARY III

SOMMAIRE IV

TABLE OF CONTENTS V

LIST OF TABLES VI

LIST OF FIGURES VI

1. INTRODUCTION 1

2. CURRENT APPLICATIONS OF VOICE INTERFACES 2

3. SPEECH OUTPUT AND VOICE DISPLAYS 4

3.1 Conditions of use for speech versus tonal or visual displays 4
3.2 Text-to-speech synthesis 5
3.3 Display types 5
3.4 Measuring system performance 6

4. SPEECH INPUT AND VOICE CONTROLS 7

4.1 Main aspects of automatic speech recognition systems 7
4.1.1 Speaker dependence *7*
4.1.2 Word boundaries *7*
4.1.3 Vocabulary size *8*
4.1.4 Measuring system performance *9*
4.1.5 System components *9*
4.2 Environmental factors affecting automatic speech recognition 10
4.2.1 Noise *10*
4.2.2 Vibration *10*
4.2.3 Stress *11*

5. SYSTEM DESIGN ISSUES 12

5.1 Speech recognition activation/deactivation 12
5.2 Multi-modal display and control 13
5.3 Feedback and error correction 13
5.4 Vocabulary size, selection and syntax 15
5.5 Robustness with noise and stressed speech 15
5.6 Overall system size and characteristics 16
5.7 Microphone and headset 16
5.8 Operating with masks and exclusion garments 16
5.9 Training 17
5.10 Security and other voice applications 17

6. CAPABILITIES OF VOICE INTERFACES 18

7. CURRENT COMMERCIAL VOICE RECOGNITION AND SYNTHESIS SOFTWARE 20

8. DISCUSSION 22

9. REFERENCES 24

List of Tables

TABLE 1: WHEN TO USE THE AUDITORY OR VISUAL PRESENTATION MODALITY 4
TABLE 2: CONDITIONS FOR USE OF VOICE INPUT (SPEECH RECOGNITION) AND OUTPUT (SPEECH SYNTHESIS).. 18
TABLE 3: FUNCTIONS FOR SPEECH RECOGNITION 19
TABLE 4: FUNCTIONS FOR SPEECH DISPLAYS 19
TABLE 5: COMMERCIAL SPEECH RECOGNITION AND TEXT-TO-SPEECH SOFTWARE 21

List of Figures

FIGURE 1: THE SOLDIER'S COMPUTER {FIGURE 3 FROM (WEINSTEIN, 1994)} 3
FIGURE 2: SPEAKING STYLE (WORD BOUNDARIES) VS. VOCABULARY SIZE {FIGURE 2 FROM (ATAL, 1994)} 8

1. Introduction

Continuous technology improvements over the years have now allowed speech to be considered as an input and output medium between soldiers and computers. Since most humans communicate with each other through speech, many consider speech as the most ‘natural’ form of interaction with computers. Speech forms the base for a natural language interface, an interface where the user can speak freely to the computer and receive a spoken response. Additionally, voice interfaces can be used when one’s hands and eyes are busy or to reduce visual workload. Unfortunately these advantages do come with limitations, such as poor speech recognition accuracy, slow performance and constrained grammar and vocabulary. These limitations can be mitigated through proper design of the computer interface for using voice in soldier tasks and environments and may ultimately be overcome by the constantly improving technology.

Voice recognition is currently being used in office environments, aids for persons with disabilities, telephone customer service and sales, voice mail, air traffic control, commercial and military airplane and helicopter cockpits, mobile phones, in-car systems, virtual reality, language training and translation, military command and control and security access (speaker identification).

This technical report investigates voice as an input and output modality for future dismounted infantry operations. In these operations soldiers are trained and armed to fight on foot and so cannot use conventional desktop input/output devices to control portable and wearable computers. The suitability of employing a voice interface in dismounted infantry operations, and the associated issues, need to be investigated as part of the Solder Information REQuirements Technology Demonstrator (SIREQ TD) project. This project investigates the effects of enhancing information capabilities in dismounted operations. This literature review will examine a Natural Language Interface, that uses voice interfaces and verbal displays to input, send, receive and display information, as a means of achieving an information exchange capability.

Current applications of voice interfaces relevant to the dismounted soldier will be discussed followed by aspects and conditions of use for speech output and voice displays. Then the main aspects of and environmental factors affecting automatic speech recognition are discussed along with issues in the design of a voice control and display system for the soldier. Current and future military capabilities of voice interfaces are noted along with a discussion of the current commercial technology available to use for laboratory and field testing.

2. Current Applications of Voice Interfaces

Various studies have been performed to investigate the application of voice displays and controls to many different domains. Those domains with lessons transferable to the dismounted soldier include wearable computing, maintenance and testing activities, robotics (cameras), air traffic control and virtual reality. Various studies are ongoing to incorporate voice into wearable computers (Farringdon & Oni, 1999; Furui, 2000; Sawhney & Schmandt, 1998; Stedmon et al., 1999). Voice interfaces have been proposed for maintenance or testing activities to allow hands-free operation at stations or while mobile (Evans, Tjoland, & Allred, 2000; G. McMillan et al., 1999). Bierschwale et al. (Bierschwale, Sampaio, Stuart, & Smith, 1989) also investigated voice input for control of camera functions during a telerobotic task. In the Military domain, Haas et al. tested automatic speech recognition for control of the radio while driving a tank in a noisy, stressful environment (Haas, Shankle, Murray, Travers, & Wheeler, 2000). Speech controls have also been investigated for use in military cockpits (Liggett, Ober, Williamson, & Reising, 1997; Nixon, Anderson et al., 1998), and helicopters (Churchill & Herdman, 2000) among others. As the technology improved, voice input/output was considered for command and control of military forces and equipment, especially when on-the-move (Ruppe & Tirabissi, 1992).

Based on visits and contacts with military organizations in the United States, Clifford Weinstein describes many opportunities of military applications of human-machine communication by voice (Weinstein, 1994). The Army could use voice communication for Command and Control on the Move (C2OTM) for forward observer reports, translations for allies, situation awareness, weapons system selection, repair and maintenance of equipment and multi-modal input/output in mobile C2 vehicles. Voice controls could also be used for radios and other auxiliary systems in helicopters. He also describes the overall concept of the Soldier's Computer (Figure 1), an Army Communications and Electronics Command program, which is very similar to the SIREQ program. Voice is considered an important input medium in these programmes since carrying and using a keyboard and other conventional input devices may be unsuitable for dismounted infantry operations.

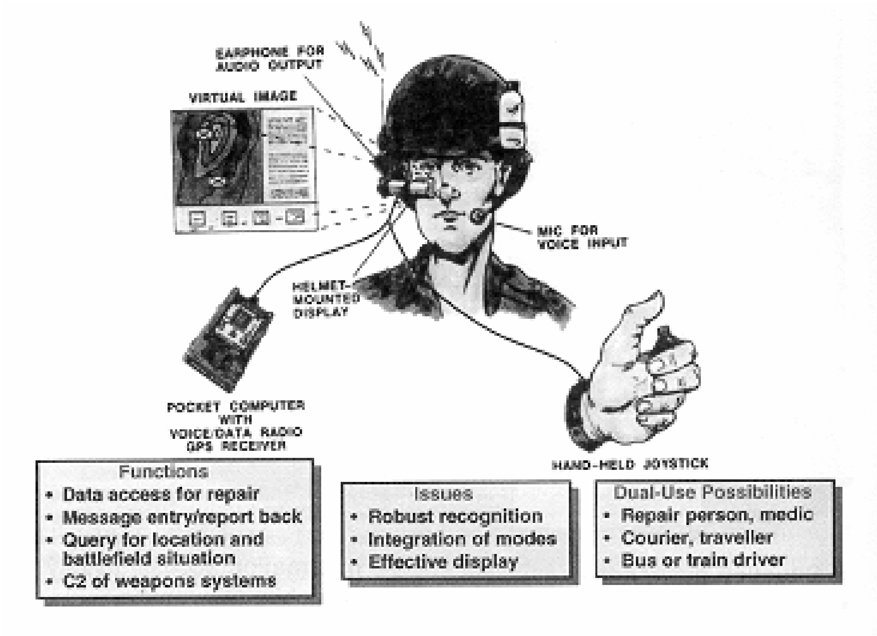

Figure 1: The Soldier's Computer {Figure 3 from (Weinstein, 1994)}

Voice control was also initially proposed for the Force XXI Land Warrior project. The Motorola Systems Solutions Group demonstrated spoken command capability to provide a nearly hands-free, eyes-free computer interface for the soldier’s computer. Recognition rates of just above 95% were obtained with environmental noise levels below 90 dB(C) (the level of hazardous noise). For very high noise (95-105 dB(C)), recognition rates were lower. Problems occurred with output commands with one syllable, five or more syllables or phonetic similarities. Performance was also unsatisfactory during yelled or very loud speech. Additionally, the noise cancelling microphone did not perform well in a substantial wind or when placed directly in front of the mouth. Even with these problems, the modified Land Warrior user interface was rated highly by the soldiers and they preferred the voice interaction to the manual command input.

3. Speech Output and Voice Displays

3.1 Conditions of use for speech versus tonal or visual displays

Speech displays use the auditory channel and have been proposed to reduce the burden on the soldier's visual channel. Speech interfaces also allow operation of equipment when the soldier's hands and eyes are busy. The conditions of use must be considered when deciding whether to use visual or auditory (speech and tonal) displays. Deatherage (Deatherage, 1972) summarizes situations when auditory and visual presentation should be used (Table 1).

Table 1: When to use the auditory or visual presentation modality

Use auditory presentation if:	Use visual presentation if:
1. The message is simple.	1. The message is complex.
2. The message is short.	2. The message is long.
3. The message will not be referred to later.	3. The message will be referred to later.
4. The message deals with events in time	4. The message deals with location in space.
5. The message calls for immediate action.	5. The message does not call for immediate action.
6. The visual system of the person is overburdened.	6. The auditory system of the person is overburdened.
7. The receiving location is too bright or dark-adaptation integrity is necessary.	7. The receiving location is too noisy.
8. The person's job requires him to move about continually.	8. The person's job allows him to remain in one position.

Source: Deatherage (1972: Table 4-1)

Using simple and short auditory messages makes sense because of the human limitations of short-term memory. If the message is long or needs to be referred to later, a "voice mail" feature for the speech interface could be used to save and playback messages. This could be particularly useful for higher-level officers such as Section or Platoon Commanders who would then be able to save incoming messages and then attend to them when they have time available. Auditory presentation is useful for warning signals that require immediate action since hearing is omni directional and cannot be ignored as easily as visual displays. Speech interfaces must be used carefully though since there is a tendency for auditory signals to pre-empt or disrupt simultaneous presentation of visual information (Cook, Cranmer, Finan, Sapeluk, & Milton, 1997). Problems could occur if the soldier needs to monitor a visual display or scene and keeps getting distracted by auditory messages.

Deatherage (Deatherage, 1972) also gives the following reasons for speech rather than tones in auditory displays:

1. Flexibility.
2. Ability to identify a message source.
3. Listeners do not need special training, as is often required for tonally coded signals.
4. Rapid two-way exchange of information is necessary.
5. Messages deal with preparation for future events. (i.e. In the countdown to firing a missile, tonal beeps could be miscounted.)
6. Situations of stress might cause the listener to "forget" the meaning of a code.

He also notes that tonal displays are preferable in situations where immediate action is desired or with high noise levels where speech becomes intelligible. Coded auditory signals can also be used to send a secure message. Tonal messages can also be used concurrently with speech messages since most tonal signals can be heard through speech and vice versa.

3.2 Text-to-speech synthesis

Speech displays depend on a text-to-speech synthesizer to create the voice prompts. Most commercial products can vary the gender, pitch, speech rate or accent of the speaker. Systems can generate speech by joining segments of pre-recorded speech or generate speech based on rules for converting letters to phonemes and phonemes to acoustic events (Kamm, 1994). The former has a more human sound to the voice, but has a limited vocabulary unlike the latter which has a theoretically infinite vocabulary. Cook also notes that human speech is easier to understand than synthetic speech, but then users assume human intelligence behind the interface and interactions may fail as a result (Cook, 1999). Some designers choose the less intelligible synthetic speech so that users are fully aware that they are talking to a computer and not a live person. Cook claims that current speech output systems lack prosody or emotional tone (Cook, 1999). Prosody adds additional meaning to the spoken text itself such as the urgency or importance of a message or the mood of the speaker. Nass and Gong (Nass & Gong, 2000), on the other hand, claim that both major methods of speech synthesis (concatenation and acoustic modeling) are capable of expressing emotional tone.

3.3 Display types

Current speech displays tend to be menu-driven or keyword-driven. Menu-driven displays are those commonly encountered with telephone interfaces where the user is presented with a choice of options and they speak or press a button for their desired choice. They continue in this structured fashion down the menu tree. Users must conform to the command structure, but this is less of a problem for soldiers than for the general public since soldiers are generally accustomed to speaking in a structured manner for radio calls. Alternatively, keyword-driven displays allow a more flexible input of command statements of single words to multiple sentences. Usually these displays use a more complicated "keyword spotting" algorithm to find the key command words. Sometimes users have difficulty remembering the options and so frequently a "What can I say?" command is included in the display, especially in the absence of a dual visual display. Future speech displays will have a natural language interface with continuous unconstrained dialogue between the operator and the computer.

Of importance especially to natural language interfaces is talk-over or barge-in capabilities. This allows the user to interrupt or begin speaking as soon as they understand the speaker's request and not wait until the speaker is finished speaking. For advanced users, this allows faster interactions but could also create problems if there is important information in the latter part of the message. Depending on the importance or time constraints of the task at hand, it will have to be decided if this capability is enabled.

To add additional information to a speech message, altering tones or words can be used. These can identify the message importance, source, type etc. Studies have been mixed as to whether this improves or slows performance. In one study, an alerting tone preceding the voice message increased response time, but an extra semantic word did not, but in a subsequent study tones used exclusively for warnings improved detection of urgent messages without increasing response time (Simpson et al., 1987). Simpson also notes that other studies of voice warning prefixes found no difference in response time as a function the prefix type of a tone, neutral word or one of three semantic cue words.

3.4 Measuring system performance

The most commonly used measure for speech displays is intelligibility. "Speech intelligibility is the percentage of utterances correctly recognized by the listener from a set of utterances presented under a given listening condition" (National Research Council, 1997). This is the same measure as used for the intelligibility of person-to-person communication and thus can be tested in the same manner with speech intelligibility tests such as the Modified Rhyme Test.

4. Speech Input and Voice Controls

4.1 Main aspects of automatic speech recognition systems

Speech input and voice controls depend on automatic speech recognition (ASR) systems. There are three main aspects when classifying ASR systems: speaker dependency, word boundaries and vocabulary size. There are also different methods to measuring system performance.

4.1.1 Speaker dependence

Speaker dependence refers to “the extent to which the system must have data about the voice characteristics of the particular human speaker(s) using it”(Simpson, McCauley, Roland, Ruth, & Williges, 1987). *Speaker-dependent* systems recognize speech from the person who trained on the system. *Speaker-independent* systems recognize speech from many speakers, not just those who trained the system. There are also speaker adaptive systems that start with speaker independent templates and then adapt those templates to the particular speaker over time. Speaker dependent systems tend to perform better than speaker-independent systems, but a significant amount of enrolment time is sometimes needed in order to train the system (can be up to 4 repetitions per word) (Sanders & McCormick, 1993, p376). Speaker-independent systems are less susceptible to large variations in speaking style and so can perform better in those situations. Additionally for the dismounted soldier, if the voice system suddenly needs to be operated by a different soldier due to injury or death, then a speaker-dependent system that cannot switch persons would not be appropriate. Since speaker-independent algorithms require more computing resources a balance needs to be achieved between system size and weight and the number of persons who can use the system with what amount of training. Ongoing developments in speech recognition software continue to expand the accuracy and power of speaker-independent systems. This will likely cease to be an issue in the next five years.

4.1.2 Word boundaries

A second classification of ASR systems is word boundaries, which refers to the amount of time that the speaker needs to place between words. In *isolated-word* systems, at least a 100 to 250 msec pause is needed between words to be recognized by the system. *Connected-word* systems need a short pause between words (< 100 msec) and no changes in intonation between words (i.e. words must be read as if from a list). *Continuous-speech* systems require no break between words and accept fluent and spontaneous speech. As the word boundary decreases, the required system resources increase. The role of the voice control system should determine the word boundaries. For example, if the soldier is using the voice control to give short commands to a piece of his equipment (eg. radio or off-bore camera) then an isolated-word or connected-word system would be appropriate. If the system is to record his verbal status report then a continuous-speech system would be appropriate. As before, a balance must be achieved between the technology and the task at hand.

4.1.3 Vocabulary size

Vocabulary size is another classification of ASR systems. A *small* system typically uses less than 200 words, a *large* system handles 1000 to 5000 words, a *very large* system handles more than 5000 words and an *unlimited* system has a vocabulary of more than 64 000 words (National Research Council, 1997). Small systems typically use voice prompts, voice commands, digit strings (telephone number dialling) and word spotting. A speech recognition system for a military aircraft cockpit had a small vocabulary of 54 words (Liggett et al., 1997). Large systems are used for applications such as dialogue with constrained semantics or directory assistance. The Defense Advanced Research Projects Agency (DARPA) Air Travel Information System (ATIS) uses a large vocabulary size as part of a transparent telephone interface (McMillan, Eggleston, & Anderson, 1997). Voice dictation systems, available commercially to use with Windows programs, use a very large vocabulary. Applications depend on both the speaking style and size of vocabulary, as shown in Figure 2.

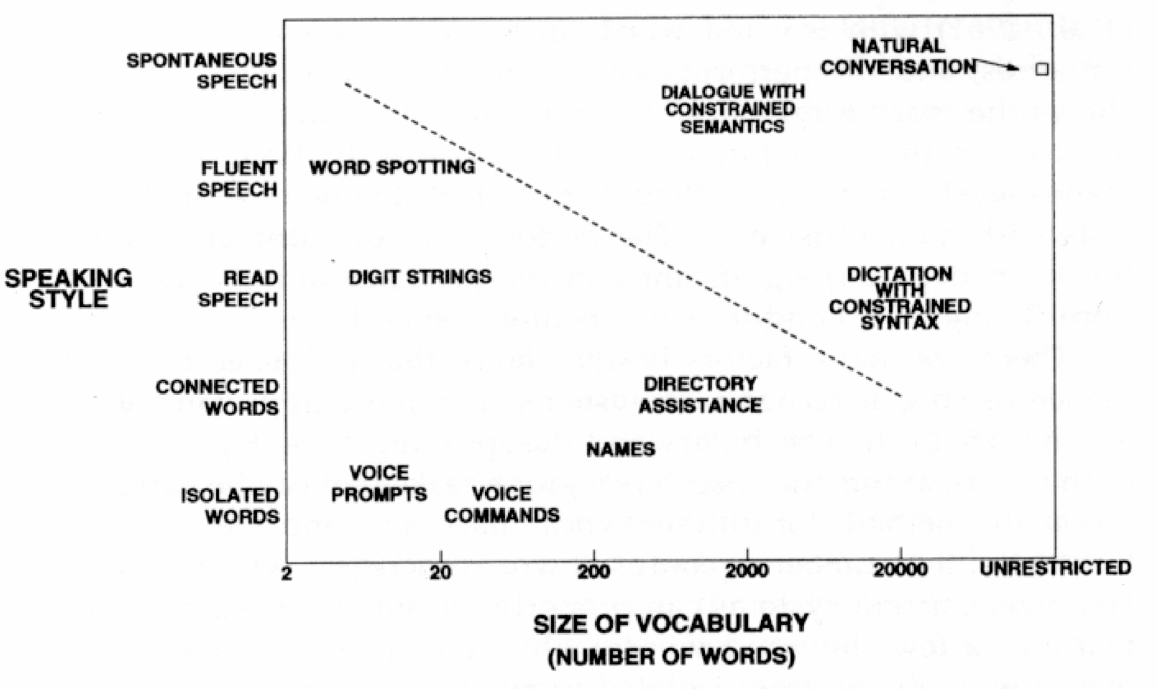

Figure 2: Speaking style (word boundaries) vs. vocabulary size {Figure 2 from (Atal, 1994)}

ASR systems can also have one or more lexicons. These ‘dictionaries’ define all the possible words in the voice system that may be encountered. Most dictation ASR systems come with a default English lexicon with entries similar to a desktop dictionary, but specialized lexicons have also been developed for the unique vocabulary found in medical, legal, public service or computer domains. ASR systems also allow the creation of custom lexicons.

As the vocabulary increases, the accuracy of the system decreases since there are more possibilities of word confusions between similar words. There is thus a balance between required vocabulary size and accuracy, though as technology improves this is less of a problem. The vocabulary should also be determined by the role that the voice control system is expected to play. It should closely match the terminology with which the soldier is familiar, while avoiding the use of acoustically similar words.

Vocabulary syntax can also be designed to allow a greater vocabulary without the loss of accuracy by enabling a sub-vocabulary at each point the command sequence. This grammar specifies which words may follow other words. *Perplexity* is the metric used to describe the complexity of a grammar and is defined as the average number of words that can follow each word in the grammar.

4.1.4 Measuring system performance

The most commonly reported statistic in early speech recognition research is *word accuracy*: the percentage of correctly recognized words in a command phrase. This statistic was applied to the performance of discrete or isolated-word systems. As connected-word and continuous speech systems became technologically feasible, phrase accuracy and intent accuracy were proposed as measures of performance (Barry, Solz, Reising, & Williamson, 1994). Since speakers were now able to enter entire command strings, *phrase accuracy* measures the percentage of complete phrases recognized correctly by the system. In order to allow flexibility of spoken commands with continuous-speech systems, the vocabulary syntax can be designed to allow multiple words to achieve the same command, such as “Display Radar” or “Show Radar”. In this case the measure of *intent accuracy*, or the percentage of correct actions performed by the system, is appropriate. Intent accuracy gives a score that does not penalize the speaker for using alternative syntax or adding extra words. The emphasis is on whether the system was able to identify the intention of the speaker’s command, regardless of how many actual words in the phrase were identified. (Liggett et al., 1997) With commands that do not have alternative forms, the intent accuracy and phrase accuracy are identical.

Liggett et al. (Liggett et al., 1997) also mention *keyword spotting* to achieve robust intent accuracy for command and control applications in a noisy environment. AT&T developed wordspotting technology in 1990 to shift the burden of speaking correctly from the user to the speech recognition algorithm for the automated telephone operator services. From their 1985 trials, they noticed that about 20 percent of user utterances contained extraneous sounds from noise and non-vocabulary words (Wilpon, 1994). Rather than force the user to speak the exact command utterance, the system was designed to listen for specific keywords associated with commands. For example, with the phrase: “I want to make a *collect* call, please”, the correctly recognized keyword *collect* will initiate a collect call.

4.1.5 System components

A typical ASR system includes a directional noise-cancelling microphone connected to an A/D converter that passes the signals into a software program consisting of a speech recognizer, a language parser and the application programs. Additionally the system may have headphones for audio feedback and a head mounted display or computer monitor for visual feedback, and other input devices such as a keyboard, touchpad, touchscreen or mouse.

4.2 Environmental factors affecting automatic speech recognition

The dismounted soldier does not operate in a quiet office environment, but rather he or she is exposed to many physical and psychological stressors that influence human speech production and the performance of speech recognition systems. Noise, vibration and stress play a major role in influencing the successful use of ASR technology.

4.2.1 Noise

High ambient noise can have a negative impact on ASR in two ways: it leads to a degradation of the speech signals but also affects the production of speech. Using noise-cancelling microphones (Broun & Campbell, 2001) or sophisticated noise cancelling and adaptation algorithms can reduce the impact of the noise on the signal. The algorithms can include “noise masking to produce a ‘clean’ speech signal and active noise compensation to provide a means of coping with fluctuations in ambient noise levels” (Baber & Noyes, 1996). High levels of noise can also lead to voice stress, a condition where people tend to increase the volume of their speech in order to hear what they are saying. This Lombard effect is a concern for the performance of ASR systems because it can lead to changes in amplitude, duration, articulation and pitch of the yelled speech as compared to quiet speech and can cause changes in the formant frequency. Various noise compensation methods have been proposed (Bou-Ghazale & Hansen, 2000; Gong, 1995). The Lombard effect can also be reduced by providing good audio feedback of the voice level in the soldier’s headset and minimizing the noise through active noise reduction techniques (Williamson, 1997). For the dismounted soldier, the ASR system must be able to accurately respond to a wide range of voice levels from low levels for stealth operations to very high levels during full combat.

Gender may be an issue with voice communications in noisy environments. Members of the Armstrong Lab at Wright-Patterson Air Force Base have done some investigations of female voice communications in aircraft noise. Their applied study (Nixon, Morris et al., 1998) indicated that the intelligibility of female speech was slightly lower than that of male speech, though not significantly. However, the intelligibility of female speech was unacceptable at the highest noise level of 115 dB. T he vulnerability of female speech to noise could be a problem with ASR systems, but they found no significant differences between the recognition accuracy of male or female speech for the ITT and IBM speech recognition systems (Nixon, Anderson et al., 1998).

4.2.2 Vibration

The dismounted soldier is exposed to vibration through vehicle operations such as helicopter and armoured personnel vehicles and also as a function of locomotion (National Research Council, 1997). These vibrations can cause “warbling” of speech sounds, breathing irregularities and tension in the jaw and thus effect speech production and the performance of ASR systems. Studies of different levels of vibration in aviation indicate that vibration may or may not be a problem (Baber & Noyes, 1996). Secondly, the “vibration” induced by shivering in extreme cold weather may also a problem in maintaining consistent speech utterances.

4.2.3 Stress

Stress accompanies military operations. The dismounted soldier can suddenly go from long periods of extreme boredom to moments of terror. These mental and physical underload and overload situations cause stress for the soldier and may affect speech production. The difficulty arises because “individuals differ in the way they are affected by environmental stressors and the way they deal with them” (Baber & Noyes, 1996). Baber notes that studies indicate that stress influences speech by altering the mean, range, variability and perturbation of the fundamental frequency and increasing the speech rate. Unfortunately, there is still a need to study the characteristics of ‘stressful speech’ so that these effects can be overcome by ASR systems to improve recognition. A NATO research group suggested that “the effect of operator based stress factors on speech production quality is likely to be detrimental to the effectiveness of communication in general, in particular to the performance of communication equipment and weapon systems equipped with vocal interfaces” (Vloeberghs, Verlinde, Swail, Steeneken, & South, 2000).

Speech research has been slow to develop in this area because of the lack of available databases of speech under stress used to test and train the proposed technology. Stressed speech cannot be recorded in actual battlefield situations and is difficult to simulate due to individual differences. The NATO research group on speech processing generated databases of speech recordings relevant to military applications (SUSC-0, SUSC-1, DLP and SUSAS) and provides them to those who are interested in using them. Commercial off-the-shelf speech recognizers where then tested using the DLP (DERA License Plate) and SUSAS (Speech Under Simulated and Actual Stress) databases. These ASR systems were not able to address the wide speaker variability associated with speech produced under stress (Vlöberghs et al., 2000). Research is continuing and there are companies that are designing their speech recognition software for noisy and/or stressful situations. ITT Industries, for example, developed speech technology (Command Voice!) for tactical applications in high noise and high stress environments.

5. System Design Issues

Designing a voice interface system for a dismounted soldier requires that the system be portable and operate in very different environments. Consideration must also be given to the activation of the system, multi-modal display and control, feedback, error correction, vocabulary, robustness, system size, and training. Other considerations, specific to a dismounted soldier, that need to be examined include the microphone and headset, operating with masks and exclusion garments, and security.

5.1 Speech recognition activation/deactivation

How and when will the speech recognition system be activated and deactivated in the field. Users must be able to quickly and easily activate and deactivate the speech recognizer (Najjar et al., 1998). For completely hands-free operation, a keyword-based system could be used but it has two drawbacks. Firstly, the keyword algorithm must run continuously, thus draining the battery and secondly, this type of system is prone to insertion errors and false command recognition. A keyword activation system was chosen for a study on direct voice input in a Griffon helicopter because the addition of a second detent or foot switch (in addition to the radio) would confuse the non-flying pilot (Churchill & Herdman, 2000). Repetition of the keyword became frustrating to the user and so for the operator interface evaluations they used an alternative method where the system was active for a period of time after the keyword was spoken and then would “go to sleep” automatically or with a command.

In noisy mobile environments it is preferable to use a push-to-talk interface so that the user can explicitly direct commands to the system or deactivate recognition completely to save system resources (Sawhney & Schmandt, 1998). For these reasons, a push-to-talk interface was chosen for the System Voice Control component of Force XXI Land Warrior (Broun & Campbell, 2001). Additionally, the location of the activation button is crucial to the soldier’s ability to easily activate the voice recognition system and so a possible solution is to place it on the soldier’s weapon (i.e. C7 rifle).

With a push-to-talk interface, it must also be decided how the system is deactivated. A button press could activate the ASR for a specified time period, as with the Land Warrior prototype for 1.5 seconds, or initiate the system to listen until a sufficiently long pause or a keyword is detected. Both of these could create problems by cutting off the end of statement prematurely, but allow hands-free deactivation. Alternatively the button could toggle the system on and off or it could be a push-and-hold-to-talk button like a tactical field radio. An advantage of the toggle button is that it could indicate the state of the ASR system, but it and the hold-to-talk button do not allow hands-free operation. The CommandTalk system uses both a push-and-hold-to-talk interface and click-to-talk that interface that listens for a pause (Moore et al., 1997).

5.2 Multi-modal display and control

For the overall system, it is recommended that multi-modal display and control should be used instead of exclusively voice since exclusively voice may not be able to be used in cases of extreme (combat) noise and stealth operations. Depending on the device, this implies having controls such as a pointing device and mini-keyboard for the soldier's computer, buttons on the radio, or buttons for an off-bore camera. Also, a visual display should be considered for both a redundant means of feedback for voice command status and for the display of other visual information for the soldier's computer. This dual functionality acts as a backup for the ASR system and can be used in cases where it may be preferable to use the manual control, such as in target acquisition.

5.3 Feedback and error correction

Feedback depends largely on the system and the task at hand and equipment available but is generally concerned with giving the system's interpretation of and response to the spoken input. It can be in a visual (textual and graphic), auditory (verbal and nonverbal) or tactile form, or be a system response such as changing a radio channel or focusing a camera. The feedback can be classified into three types: *reactive*, *instrumental* or *operational* (Baber, 2001). *Reactive* feedback occurs when using a control, for example when a button is pressed it moves and clicks. There is usually no reactive feedback when using a microphone, and so the speaker may not know if the mic is "live", whether they are speaking loudly enough, or whether the signal is reaching the speech recognizer. Early isolated-word speech recognition systems used a beep to indicate that the system is ready, but this irritated advanced users. There is no current solution for effective reactive feedback with ASR systems. Secondly, the performance of the speech recognizer itself is shown through *instrumental* feedback, which tells the user what words have been recognized. It can be concurrent (after each word, digit, letter, etc.) or terminal (after a command, phrase, digit string, etc.). Both have limitations since concurrent feedback may disrupt command utterances (high working memory load) but error handling is more difficult with terminal feedback. Finally, *Operational* feedback involves the system responses to spoken commands, such as changing views on a visual display or switching a radio frequency.

Appropriate feedback is critical to the overall performance and user acceptance of the system. Without feedback, the user may incorrectly assume that a sequence of voice commands was executed by a system (McMillan et al., 1997). Any communications with the user should be simple and immediate (Cleveland & McNinch, 1999). For example, when the user's speech pattern is not acceptable, the system should express the feedback in simple language such as "Speak Louder", through an auditory icon, or a short text message or graphic on a screen. Feedback should not be simply repeating or displaying every entry that is recognized by the system, but rather give feedback for events, system status, unrecognized inputs or user interface errors. The feedback should provide a sense of control and assure the user that the system is working while not hindering or annoying users (Najjar et al., 1998).

For the dismounted soldier, feedback could be achieved in several ways. Visual feedback can be provided on a head-mounted display, portable display or a digital display located on the weapon or other equipment. If a small display is available to the soldier, Schurik et al. (1985) recommend giving visual instrumental feedback concurrently (by word) or terminally (by field).

Conversely if the visual display is not available or may overload the visual channel, they recommend the slower auditory concurrent instrumental feedback. Auditory feedback could improve error detection, as Noyes and Frankish (1994) found that subjects monitor spoken feedback better than visual feedback. Provided the soldier has a headset, speech synthesis can be used to provide verbal feedback. For example, in a hands-free radio dialogue the soldier could say “Goto Foxtrot now”, and instead of him manually changing the dial on the radio, the ASR system recognizes the command and changes the channel. He hears the operational feedback: “Radio set to Foxtrot” and then does a radio check to confirm the correct frequency. Alternatively, mono or 3D sounds could be used as auditory icons to give feedback for mode transitions, system status, errors and accepted commands. Tactile signals could also be used in place of auditory icons for cases where the auditory channel is needed for a concurrent task. Some of the feedback should be concurrent (such as system commands) but other feedback could be terminal (proofreading a verbally entered status report). Further investigation is needed to determine the appropriate combination of feedback modalities.

Some of the feedback leads to error correction of incorrectly recognized speech. Errors can be false insertions (the speech engine inserts extra words), substitutions (the engine replaces one word with another) or deletions (the engine drops a word). The simplest form of error correction is to have the user repeat the last utterance. Speakers prefer to use this method rather than using the system’s error correction techniques (Najjar et al., 1998) such as word suggestion based on the output of a pattern-matching algorithm. Ainsworth and Pratt (1992) found that the error-correction strategy of ‘repetition-with-elimination’ required fewer trials to correct the recognition errors than ‘elimination-without-repetition’. In the former strategy, when the user detects a mistake, the system eliminates the last response from the active vocabulary and then the user repeats the command. Eventually with a finite vocabulary, the correct response will occur. For the latter ‘elimination-without-repetition’ strategy, the system suggests the next most likely word when an error is detected. Murray et al. (1993) used similar strategies to Ainsworth and Pratt but produced a different conclusion. For a highly confusable vocabulary, their choice procedure (elimination-without-repetition) was the most efficient in terms of the number of successful corrections and the speed with which they were completed. These different methods of feedback will have to be considered and determined in the design of the task for the dismounted soldier.

Multi-modal error correction is also possible if the soldier has a pointing or input device for the computer, or manual controls on the equipment. For example, the radio channel could be changed manually or the soldier could use the pointing device to select incorrect text and then speak the correct text. This backup form of error correction is very useful when the performance of the ASR system degrades due to environmental factors. Additionally, the system could provide different methods of error correction depending on whether the error occurred with a command or with dictated text. Overall, the speech recognition interface needs to be designed so that error correction is simple and obvious.

5.4 Vocabulary size, selection and syntax

When designing the vocabulary, the command phonetics need to be distinctly different from each other in order to avoid short commands with similar vowel sounds that will be confused by the recognizer. The vocabulary should also be task dependent (i.e. radio calls for communication equipment) and kept as small as possible to avoid word confusions between similar words. A vocabulary syntax, or lexicon, should also be used to limit the subset of the vocabulary available at any time. Based on the command context, the language parser rejects those recognized words not in the available vocabulary. This form of out-of-vocabulary rejection helps to reduce insertion and substitution errors.

As an example of vocabulary syntax, SRI International developed an application-specific grammar for the CommandTalk spoken language dialogue system. CommandTalk is a spoken-language interface to synthetic forces in entity-based battlefield simulations. It lets simulation operators interact with synthetic forces by voice in a manner as similar as possible to the way that commanders control live forces. Examples of the grammar include:

> (MOVEMENT_TYPE to POINT_OR_LINE and engage ENEMY_LOC with direct fire)
>
> (rendezvous with UNIT_CALL_SIGN at POINT_LOC)

Each of the capitalized words has sub-rules associated with it, such as UNIT_CALL_SIGN is composed of (ICA_LETTER DIGIT DIGIT) or (ICA_LETTER DIGIT DIGIT DIGIT) and then ICA_LETTER consists of (alpha, bravo, charlie, etc.). There are also system commands such as (zoom in on UNIT) or (center COMPASS_DIRECTION of [UNIT POINT_LOC]). Referring to the example above, if the operator has spoken "rendezvous with", then the engine will attempt to match the spoken input to speech templates for ICA letters and reject other inputs.

Specialized vocabulary syntax also requires less processing resources. CommandTalk uses the Nuance speech recognition system and the Gemini natural-language parsing and interpretation system. By setting Gemini to use only specific military vocabulary and grammar, the grammar parsing runs faster with fewer ambiguities and less processing than the extensive grammar of general English (Moore et al., 1997). This performance increase is advantageous when applying voice recognition to mobile computing.

5.5 Robustness with noise and stressed speech

The soldier environment is very demanding and so the voice control system must be robust to noise and stressed speech. These problems are still being researched to achieve better performance. Vlöberghs et al. (2000) describe in detail stress compensation techniques to achieve robust speech recognition through better training methods, improved front-end processing, and improved back-end processing or robust recognition measures. Chapter 6 of their report gives details of the specific measures investigated. Other research is also being done to improve voice recognition with stressed speech (Bou-Ghazale & Hansen, 2000) and in noisy environments (Singh, Seltzer, Raj, & Stern, 2001), among others. A survey of research in noisy speech recognition is provided in (Gong, 1995).

5.6 Overall system size and characteristics

The equipment needed for the voice control system will ultimately add weight (and cost) to the soldier’s pack and so thus needs to be as small and light as practically possible. The algorithms of the ASR system must also be computationally efficient as to not drain the system battery, and the software must be sufficiently small to fit in the available memory on the soldier’s computer (Broun & Campbell, 2001). For example, the Motorolla CVoxCon speech recognition engine with polynomial classifiers was selected for the Land Warrior demonstration because this low complexity technology is suitable for portable devices (Campbell, Assaleh, & Broun, 1999).

External components, such as a headset and microphone, must also be compatible with current clothing and equipment. All components must be very durable and robust to handle field use and designed such that if one component fails or is damaged, the whole system does not crash. The system should also be designed so that another soldier can take over the system quickly and easily, should the current user be incapacitated (Stedmon et al., 1999).

5.7 Microphone and headset

The voice controls require a microphone and the voice displays require a headset or speaker. For operating in noisy environments, it is critical to have the microphone placed in a fixed optimum location. Highly directional, noise cancelling traditional or throat/bone conducting microphones also help to reduce the effect of background noise. Najjar et al. (1998) were able to successfully test a speech-driven system with this type of microphone in a very loud 90dB noise environment. In a related study, the traditional noise-cancelling microphone used for a Land Warrior prototype did not perform well in a substantial wind or when placed directly in front of the mouth (Cleveland & McNinch, 1999).

Consideration should be given to a headset or earphone that is noise-immune such as ear canal or bone conduction earphones. This listening apparatus would also be used with other communications equipment and other auditory tasks. Monaural (one ear) or binaural (both ears) can be used depending on the other communications tasks. For monaural presentation, the right ear is preferred for speech (left-brain dominance) and the left ear is preferred for music and other sounds (Nass & Gong, 2000).

Overall the microphone and headset must not interfere with weapon use and must operate in all weather conditions including rain, snow and wind. Additionally, they must be compatible with the soldier’s clothing and equipment, including helmet, cap, cold weather mask, NBC C4 mask, comms (communications equipment), eyewear and any coats or vests that cover parts of the face, neck or head.

5.8 Operating with masks and exclusion garments

The NRC panel noted that “voice entry while wearing some form of mask or exclusion garment may prevent optimal use of critical systems” (National Research Council, 1997). Not only may the clothing or equipment physically interfere with the microphone and headset, but also the performance of the ASR system could be reduced if the mask changes the soldier’s speaking tone and rate. If the speaking style does significantly change, then the soldier should also train speech templates for the ASR while wearing the mask. When the garment or mask is then worn out in the field, the ASR system then switches manually or automatically to the “mask” templates.

5.9 Training

There are two aspects to training: training a speaker-dependent ASR system to the speaker’s voice and training the soldier to use the voice input/output system. If possible, train a speaker-dependent system in an environment similar to where it will be used (Najjar et al., 1998). This significantly improves the recognition accuracy by having the user speak in a “style similar to that when working” and allows the recognition to adapt to the ambient noise. The soldier ASR system may not be able to depend on this general recommendation for training though since the speaking style and noise level varies dramatically for the dismounted soldier. The ASR system cannot be trained during actual combat to sample noise levels. As for soldier training, the soldier needs to learn the command vocabulary of the system and what speaking style gives the best accuracy. If an isolated-word system is in use, the speaker must also learn to insert the pause between words. He or she must additionally become familiar with the structure and functioning of the voice menu commands and displays.

5.10 Security and other voice applications

With any military equipment, the question arises as to the consequences if hostile forces capture it. Will they be able to use it and/or gain tactical information from it? Speaker-dependent systems can hinder use by those not trained on the system, but an additional capability of speech technology is speaker recognition. This involves speaker verification, speaker identification and language identification (Weinstein, 1991). The speaker’s voice can act as an auditory password, thereby enabling the system to determine if the user is authorized to use the system. Unfortunately though, if the voice identification system is not robust there could be serious consequences if the system does not recognise the true soldier’s voice due to illness, stress or other environmental factors. Speaker identification can also be used for surveillance of communications channels to identify those talking.

Language identification would be useful for the Canadian Forces since a voice recognition system would then be capable of determining if the speaker is speaking French or English and then use the appropriate speech recognition algorithm and employ the appropriate language text in any visual and auditory displays.

It is also being used in automatic interpretation to allow persons to speak to each other in different languages. For example, the INTERTALKER system recognizes Japanese and English speech and translates and synthesizes it into English, Japanese, French and Spanish (Kato, 1994). Automatic language translation would solve many communications problems when Canadian soldiers are interacting with non-English/French speaking coalition forces and local foreign citizenry.

6. Capabilities of Voice Interfaces

The capabilities of voice interfaces depend on where and when the system is used and what tasks are to be performed using the system. The basic conditions or situations of use for the soldier (related to voice interfaces) are shown in Table 2 along with the suitability of use and further comments. The following interpretations were made: “no” for not possible at this time, “possible” for being achievable with some limitations, “yes” for being achievable, and “?” for being unsure. This table can be used as a reference when determining the conditions under which voice input and output should be used for the dismounted soldier.

Table 2: Conditions for use of voice input (speech recognition) and output (speech synthesis)

Condition	Voice Input	Voice Output	Comments
Constant ambient noise (up to 120 dB)	Possible to yes	Possible to no	For voice input, some ASR systems can be calibrated or adjusted for high noise levels. Speech output could become unintelligible if the wearer is not insulated from high ambient noise levels (eg. earphone, earplug).
Noise bursts (i.e. tank fire)	No	No	Speech is no longer intelligible.
Operating at night	Yes	Yes	Performance is independent of the amount of daylight. However, other backup modalities (i.e. a visual display for feedback) may not be independent.
Walking	Yes?	Yes	Does marching noise interfere with the speech recognizer?
Running	Possible?	Yes	Breath noises may interfere with the recognizer. Speaking style may change and lower performance. The microphone may vibrate and effect speech capture quality.
Any prone position (lying down, standing, crouching, etc.)	Yes	Yes	Recognition accuracy may decrease as the soldier’s stress level increases.
Hiding and stealth operations	Possible to no	Possible	Whispered voice has poorer recognition. The enemy could detect speech sounds.
Yelling	Possible	N/A	Can lead to Lombard speech, which has poorer recognition. Stress compensation algorithms can be used to improve recognition.
Wearing a mask	Possible	?	The mask must accommodate the microphone and headset. Changes in the speaker’s voice and breath noises decrease ASR performance so special speech templates and/or compensating algorithms may need to be used.
Completely hands-free	Possible	Yes	Hands-free voice input uses more system resources and increases insertion errors compared to a push-to-talk interface.
Very high-stress situations	?	Possible	Voice output should be used sparingly since it is slow and can interfere with visual processing. Voice input needs to be investigated for specific situations since it can be used to allow quick hands-free input, but under stress soldier may forget the command structure and/or change their speaking style.

Voice interfaces also have certain capabilities that influence the tasks that can be effectively performed with voice input. The suitability of using speech recognition (voice controls) is summarized for select infantry functions in Table 3, along with the suitability of speech displays in Table 4. Specific tasks will need to be investigated in more detail through readings or laboratory testing.

Table 3: Functions for speech recognition

Function	Use	Comments
Data entry	Yes	Long texts are easier with a continuous speech ASR system. An error correction method based on the available input technology would need to be designed. Type of feedback needs to be determined.
Issuing voice commands to equipment or indirectly to personnel	Yes	Recognition improves with use of vocabulary syntax.
Positioning or moving objects	Possible to no	Use a pointing device if the location for the object cannot be uniquely identified. Using speech is slower and less precise than manual input (Bierschwale et al., 1989) and annoys users (Najjar, Ockerman, & Thompson, 1998).
Navigating through software	Possible	Screens and functions can be uniquely defined by a voice command. Options may be difficult to recall from memory. Type of feedback needs to be determined.
Language translation	Possible in future	Allows multi-lingual communication, but technology needs to develop further.
Speaker identification	Possible	Need supporting software.

Table 4: Functions for speech displays

Function	Use	Comments
Warnings	Yes	Very good at attracting attention. Should be worded as short phrases with four or five syllables. Can have preceding alert tones or words. Use a distinctive voice to not be confused with other human voices.
Advisories	Possible	Can be confused with warnings.
Prompts	Yes	There is a balance between verbosity (task completion time) and errors.
Feedback	Yes	Different types of feedback (short/long verbal, beeps/tones, visual, etc.) should be used depending on the task and time available.
Responses to user queries	Possible	Avoid very long texts that would be difficult to hold in short term memory. This speech function avoids users from having to divert their visual attention to search a display or device for the desired information (i.e. current radio channel or bearing)
Commands (to the user)	?	Simpson (1985) notes that pilots were reluctant to follow a command (by an automatic speech generator) without knowing the reason for it. Could be useful for instructions in situations that are not time-critical.

(adapted from Simpson, 1985)

7. Current Commercial Voice Recognition and Synthesis Software

Further laboratory and field investigations of voice controls and displays require the selection of suitable voice recognition and synthesis software. This software would not necessarily be used in the final soldier system but would be used in any experimentation platform. The aim is to find a product that can demonstrate the capability of the voice modality with the least amount of development time and costs.

The main providers for voice recognition software for desktop PCs are IBM Corporation (*ViaVoice*) and Scansoft (*Dragon NaturallySpeaking*). These continuous speech speaker independent systems are designed mostly for dictation and web browsing. The text-to-speech technology provides speech synthesis for documents and dictated text. To improve recognition accuracy, they can be trained for the specific speaker. Because of the large vocabulary (260 000 and 250 000 words), they require at least a PII 300 MHz processor with 510 or 300 MB free disk space and 64-96 or 128-256 MB RAM respectively. They are not primarily designed for high noise environments, but can be used if trained in the noise. For custom applications, speech interfaces can be designed using the SDKs, ActiveX controls or directly interfaced to the SMAPI (IBM) or SAPI speech engines. This type of software could be used for a demonstration prototype, but the processing and space requirements would place a large burden on wearable computing resources.

A second area of voice recognition software is in embedded devices such as mobile phones, PDAs, toys and kitchen appliances. Philips (*VoCon* and *SpeechWave VR*), IBM (*Embedded ViaVoice*), 20/20 Speech (*Aurix asr* and *tts*) and Conversay (*Mobile Conversay*) are some of the major providers. The software for embedded devices tends to be speaker independent, provides text-to-speech synthesis, has a small vocabulary for command and control, and requires very little processing and disk space. Additionally, *Mobile Conversay* and *VoCon* are noise robust. *Aurix asr* was initially developed for military purposes by the former Defence Evaluation and Research Agency (DERA) in the United Kingdom. Product-specific software development kits (SDKs) are available to create custom applications. The processing requirements meet the specifications for a wearable computer, but some software may not operate on the soldier's computer platform. Additionally, tests would have to be performed to see if it is noise robust for the soldier domain.

The remainder of voice recognition software is focused on web, network and telephony services such as customer support and directory assistance. Web servers are not appropriate for the soldier's computer, but a related product from Conversay, Voice Surfer, has been used for voice recognition for the Xybernaut wearable computer. This continuous speech speaker-independent voice-browser acts as an extension to Internet Explorer 5.5 and requires very little disk space (15 MB) and processing speed (PI 200 MHz). Interfaces can be designed using the Voice Surfer Development Kit and Macromedia Dreamweaver 4. This product could be used to create a simple interface to the soldier system, but it is dependent on IE 5.5. Telephony services are also not applicable to the soldier domain, but the core processor of the Nuance system has been used for speech recognition in CommandTalk (Moore et al., 1997) and ITT Industries'

CommandVoice! Tactical Voice Recognition System (TVRS). CommandVoice! was designed primarily for military and commercial industrial applications with high noise and stress. The ITT-1290 Speech Recognition system was tested in military aircraft cockpits and obtained 95.2% average intent accuracy in level flight (Liggett et al., 1997). This speaker-dependent continuous speech system provides automatic gain control, speaker channel normalization, speech synthesis and noise calibration. The speech templates are trained quickly and the vocabulary syntax can be switched dynamically. The software runs on Windows NT/2000, Linux or Solaris and needs only 5-10 MB RAM with 15 MB hard disk space plus 1 MB per user for the voice models.

The products and contact information are summarized in Table 5 on the next page. For the SIREQ TD experiments, the software by 20/20 Speech and ITT Industries look the most promising. Both have been used or tested in the military domain and have low processing requirements and noise adaptation. The main difference is that Aurix asr is speaker-independent whereas CommandVoice is speaker-dependent. System performance, compatibility with the trial computers (Xybernauts), development time and cost will be the main factors in the choice of software to use for the laboratory or field trials.

Table 5: Commercial Speech Recognition and Text-to-Speech Software

Company	Product	Contact
20/20 Speech	Aurix asr & tts	http://www.2020speech.co.uk/
Conversay	Mobile Conversay Voice Surfer	http://www.conversay.com/Products/Embedded/default.asp http://www.conversay.com/Products/Browsers/default.asp
IBM Corp.	ViaVoice 9.0 Embedded ViaVoice	http://www-3.ibm.com/software/speech/desktop/w9.html http://www-3.ibm.com/pvc/products/voice/ vv_mobile_device.shtml
ITT Industries Aerospace/ Communications Division	CommandVoice! Tactical Voice Recognition System (TVRS)	Mark Bradley mark.bradley@itt.com (219) 451-7788 http://www.acd.itt.com/c_voice.htm
Nuance	Nuance 8.0 Vocalizer 2.0	http://www.nuance.com/prodserv/prodnuance.html http://www.nuance.com/prodserv/prodvocalizer.html
Philips	VoCon SpeechWave VR	http://www.speech.philips.com/vc/Pages/vc_home.htm
Scansoft	Dragon NaturallySpeaking 6.0	http://www.scansoft.com/naturallyspeaking/

8. Discussion

Voice controls and displays can be an effective information control and display interface for the dismounted soldier but the limitations and capabilities must be taken into consideration. This input/output modality currently tests the limits of the technology, but does allow computer system operation when the soldier's hands and eyes are busy, a potentially critical benefit.

Voice displays are best used when the soldier is moving and the operational conditions do not allow the use of a visual display. Speech presentation is best for short, simple messages that do not need to be referred to later in time. Voice displays require a text-to-speech synthesizer to produce the speech messages through phonemes or recorded text. Users prefer when the synthesized voice is distinguishable from other human voices. Current displays are usually menu or keyword driven, but as the technology improves, a natural language interface is possible. Altering tones or cue words, preceding the voice message, can be used to add additional information to a speech message. Speech displays are well suited to warnings, prompts, feedback, and short responses to user queries. Advisories should be used with caution since they can be confused with warnings.

Automatic speech recognition (ASR) systems are the basis behind voice controls. The main characteristics of these systems are the dependency on the speaker, word boundaries and vocabulary size. As each of these becomes more complicated, the system size and cost increases and word or phrase accuracy may decrease because of word confusions. The performance of the ASR system is worsened by environmental factors such as noise, vibration and stress. These effects can be reduced through compensating algorithms and human factors aspects in the design of the system. Speech recognition is best suited for data entry, issuing voice commands, language translation and speaker identification. It should not be used for moving or positioning objects in a visual display.

There are also several issues in the general design of the system. Speech recognition systems can be continuously active or employ some form of a push-to-talk interface. A multi-modal display and control system should be employed to provide a backup to the voice system in cases where voice may not be appropriate and in situations where redundant feedback is desirable. The system also needs to provide some form of feedback to the soldier so that error correction is simple and obvious.

Based on these capabilities and limitations, a voice interface system for dismounted soldiers should consider the following design issues.

Voice Displays:

The voice display should be based on simple or short messages that require immediate attention. Avoid the use of too many voice messages when the soldier's visual and/or auditory systems are overburdened. A 'save' function should be provided to the soldier if the message needs to be referred to later in time. To distinguish the voice display from the surrounding voices in the soldier's environment, use either a synthetic 'computer voice' or the opposite gender for the speech synthesis. Laboratory tests should be conducted to decide if an alerting tone or word should be used to identify a speech message.

Voice Controls:

A robust noise-immune automatic speech recognition system should be chosen for the soldier's computer. Ideally it should be speaker independent, continuous-speech with a low perplexity task-dependent grammar. To decrease word confusions, commands need to be phonetically different from each other in both the basic phonemes (sounds) and/or the number of syllables. Out in the field, the system needs to be able to quickly adapt to whispered speech, yelled speech, changes in background noise, and talking while wearing a mask. A form of push-to-talk interface should be used to avoid insertion errors and preserve system resources. The type of feedback and error correction needs to be determined in further tests. The system should operate with a word or phrase accuracy of 95% for data entry and 99% for command and control.

System Design:

In addition to speech recognition and synthesis software, a microphone and headset are also required. These should be noise-immune such as directional or throat microphones and ear canal or bone conduction earphones. Given the general need to minimize size and weight requirements on soldiers, all equipment should be as small as possible, robust to operate in all types of environments and compatible with the soldier's clothing and equipment. For a prototype system to use in laboratory or field trials, the software available from 20/20 Speech or ITT Industries appear most promising.

9. References

AINSWORTH, W.A., & PRATT, S.R., 1992. Feedback Strategies for Error Correction in Speech Recognition Systems. *International Journal of Man-Machine Studies, 36*(6), 833-842.

BABER, C., 2001. Interactive Speech Technology. In W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors* (Vol. 1, pp. 698-700). London: Taylor and Francis.

BABER, C., & NOYES, J., 1996. Automatic Speech Recognition in Adverse Environments. *Human Factors, 38*(1), 142-155.

BARRY, T., SOLZ, T., REISING, J., & WILLIAMSON, D., 1994. The use of word, phrase, and intent accuracy as measures of connected speech recognition performance. *Paper presented at the Human Factors and Ergonomics Society 38th Annual Meeting*. October 24-28, Nashville, TN, USA.

BIERSCHWALE, J.M., SAMPAIO, C.E., STUART, M.A., & SMITH, R.L., 1989. Speech versus Manual Control of Camera Functions during a Telerobotic Task. *Paper presented at the Perspectives: The Human Factors Society 33rd Annual Meeting*, October 16-20, Denver, Colorado.

BOU-GHAZALE, S.E., & HANSEN, J.H.L., 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing, 8*(4), 429-442.

BROUN, C.C., & CAMPBELL, W.M., 2001. Force XXI Land Warrior: a systems approach to speech recognition. *Paper presented at the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 7-11, Salt Lake City, UT, USA.

CAMPBELL, W.M., ASSALEH, K.T., & BROUN, C.C., 1999. Low-Complexity Small-Vocabulary Speech Recognition for Portable Devices. *Paper presented at the Fifth International Symposium on Signal Processing and its Applications, ISSPA '99,* August 22-25*,* Brisbane, Australia.

CHURCHILL, L.L., & HERDMAN, C., 2000. CH146 Griffon direct voice input control for the Canadian Forces utility tactical transport helicopter: Final project report. *BAE Systems Canada Inc.*, Kanata, Ontario.

CLEVELAND, G., & MCNINCH, L., 1999. Force XXI Land Warrior: Implementing Spoken Commands for Soldier Wearable Systems. *Paper presented at the Third International Symposium on Wearable Computers*, October 18-19, San Francisco, California.

COOK, M., 1999. Speech I/O. In J. M. Noyes & M. Cook (Eds.), *Interface Technology: The Leading Edge* (pp. 45-58). Baldock, Hertfordshire: Research Studies Press.

COOK, M., Cranmer, C., Finan, R., Sapeluk, A., & Milton, C., 1997. Memory Load and Task Interference: Hidden Usability Issues in Speech Interfaces. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics*. (Vol. 1 - Transportation systems, pp. 141-150). Aldershot, Hampshire: Ashgate Publishing.

DEATHERAGE, B.H., 1972. Auditory and other sensory forms of information presentation. In H. P. Van Cott & R. G. Kinkade (Eds.), *Human engineering guide to equipment design* (Revised ed., pp. 123-160). Washington, DC: US Government Printing Office/McGraw Hill Company.

EVANS, J.R., TJOLAND, W.A., & ALLRED, L.G., 2000. Achieving a hands-free computer interface using voice recognition and speech synthesis [for Windows-based ATE]. *IEEE Aerospace and Electronics Systems Magazine, 15*(1), 14-16.

FARRINGDON, J., & ONI, V., 1999. Co-Modal Browser - An Interface for Wearable Computers. *Paper presented at the Third International Symposium on Wearable Computers,* October 18-19, San Francisco, CA, USA.

FURUI, S., 2000. Speech recognition technology in the ubiquitous/wearable computing environment. *Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00,* June 5-9, Istanbul, Turkey.

GARDNER-BONNEAU, G. (Ed.)., 1999. Human Factors and Voice Interactive Systems. Boston: Kluwer Academic Publishers.

GONG, Y., 1995. Speech Recognition in Noisy Environments: A Survey. *Speech Communication, 16*(3), 261-291.

HAAS, E., SHANKLE, R., MURRAY, H., TRAVERS, D., & WHEELER, T., 2000. Issues Relating to Automatic Speech Recognition and Spatial Auditory Displays in High Noise, Stressful Tank Environments. *Paper presented at Ergonomics for the New Millennium: The XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society*, San Diego, California, USA, July 29 - August 4.

KAMM, C., 1994. User Interfaces for Voice Applications. In D. B. Roe & J. G. Wilpon (Eds.), *Voice Communication between Humans and Machines* (pp. 422-442). Washington, D.C.: National Academy Press.

KATO, Y., 1994. The Future of Voice-Processing Technology in the World of Computers and Communications. In D. B. Roe & J. G. Wilpon (Eds.), *Voice Communication Between Humans and Machines* (pp. 505-514). Washington, D.C.: National Academy of Sciences.

LIGGETT, K. K., OBER, K. R., WILLIAMSON, D. T., & REISING, J. M., 1997. *Speech Recognition Systems for Military Aircraft Cockpits: From Laboratory to Flight Test.* Paper presented at the Proceedings of the Ninth International Symposium on Aviation Psychology, April 27 - May 1, Columbus, OH.

MCMILLAN, G., CALHOUN, G., MASQUELIER, B.L., GRIGSBY, S.S., QUILL, L.L., KANCLER, D.E., & REVELS, A.R., 1999. *Comparison of Hands-Free versus Conventional Wearable Computer Control for Maintenance Applications.* Paper presented at Houston… We Have a Solution!: The Human Factors and Ergonomics Society 43rd Annual Meeting, September 27 - October 1, Houston, Texas.

MCMILLAN, G.R., EGGLESTON, R.G., & ANDERSON, T.R., 1997. Nonconventional controls. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (Second ed., pp. 729-771). New York, NY: John Wiley & Sons, Inc.

MOORE, R., DOWDING, J., BRATT, H., GAWRON, J.M., GORFU, Y., & CHEYER, A., 1997. CommandTalk: A Spoken-Language Interface for Battlefield Simulations. *Paper presented at the Fifth Conference on Applied Natural Language Processing,* Washington, DC.

MURRAY, A.C., FRANKISH, C.R., & JONES, D.M., 1993. Data-entry by voice: facilitating correction of misrecognitions. In C. Baber & J. M. Noyes (Eds.), *Interactive Speech Technology: Human factors issues in the application of speech input/output to computers*. London, UK: Taylor & Francis Ltd.

NAJJAR, L.J., OCKERMAN, J.J., & THOMPSON, J.C., 1998. *User Interface Design Guidelines for Wearable Computer Speech Recognition Appliances.* Paper presented at the IEEE VRAIS 98 Workshop - Interfaces for Wearable Computers.

NASS, C., & GONG, L., 2000. Speech Interfaces from an Evolutionary Perspective. *Communications of the ACM, 43*(9), 36-43.

NATIONAL RESEARCH COUNCIL, 1997. Tactical Display for Soldiers: Human Factors Considerations. Washington, D.C.: National Academy Press.

NIXON, C.W., ANDERSON, T.R., MORRIS, L.J., MCCAVITT, A.R., MCKINLEY, R.L., YEAGER, D.G., & MCDANIEL, M.P., 1998. Female voice communications in high level aircraft cockpit noises: Part II. Vocoder and automatic speech recognition systems. *Aviation, Space, and Environmental Medicine, 69*(11), 1087-1094.

NIXON, C.W., MORRIS, L.J., MCCAVITT, A.R., MCKINLEY, R.L., ANDERSON, T.R., MCDANIEL, M.P., & YEAGER, D.G., 1998. Female voice communications in high levels of aircraft noises. I - Spectra, levels, and microphones. *Aviation, Space, and Environmental Medicine, 69*(7), 675-683.

NOYES, J.M., & FRANKISH, C.R., 1994. Errors and error correction in automatic speech recognition systems. *Ergonomics, 37*(11), 1943-1957.

SANDERS, M.S., & MCCORMICK, E.J., 1993. Human Factors in Engineering and Design (Seventh ed.). New York: McGraw-Hill, Inc.

SAWHNEY, N., & SCHMANDT, C., 1998. Speaking and Listening on the Run: Design for Wearable Audio Computing. *Paper presented at the Second International Symposium on Wearable Computers, Digest of Papers*, October 19-20, Pittsburgh, Pennsylvania.

SCHURICK, J.M., WILLIGES, B.H., ET AL., 1985. User feedback requirements with automatic speech recognition. *Ergonomics, 28*(11), 1543-1555.

SIMPSON, C.A., MCCAULEY, M.E., ROLAND, E.F., RUTH, J.C., & WILLIGES, B.H., 1987. Speech Controls and Displays. In G. Salvendy (Ed.), *Handbook of Human Factors* (pp. 1490-1525). New York: John Wiley & Sons, Inc.

SINGH, R., SELTZER, M.L., RAJ, B., & STERN, R.M., 2001. Speech in Noisy Environments: robust automatic segmentation, feature extraction, and hypothesis combination. *Paper presented at the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 7-11, Salt Lake City, UT, USA.

STEDMON, A.W., KALAWSKY, R.S., MOORE, P.M., AUNG, M., PURCELL, C., REEH, C., & YORK, T., 1999. It's not what you wear, it's how you wear it: human factors of wearable computers. *Paper presented at Houston… We Have a Solution!: The Human Factors and Ergonomics Society 43rd Annual Meeting*, September 27 - October 1, Houston, Texas.

VLOEBERGHS, C., VERLINDE, P., SWAIL, C., STEENEKEN, H., & SOUTH, A., 2000. *The Impact of Speech Under "Stress" on Military Speech Technology* (RTO-TR-10 AC/323(IST)TP/5). Neuilly-sur-Seine (France): NATO Research and Technology Organization.

WEINSTEIN, C.J., 1991. Opportunities for Advanced Speech Processing in Military Computer-Based Systems. *Proceedings of the IEEE, 79*(11), 1626-1641.

WEINSTEIN, C.J., 1994. Military and Government Applications of Human-Machine Communication by Voice. In D. B. Roe & J. G. Wilpon (Eds.), *Voice Communication between Humans and Machines* (pp. 357-370). Washington, D.C.: National Academy Press.

WILLIAMSON, D.T., 1997. Robust speech recognition interface to electronic crewmember: progress and challenges. *Paper presented at the Proceedings of 4th Human-Electronic Crewmember Workshop*, Kreuth, Germany.

WILPON, J.G., 1994. Applications of Voice-Processing Technology in Telecommunications. In D.B. Roe & J.G. Wilpon (Eds.), *Voice Communication Between Humans and Machines* (pp. 280-310). Washington, D.C.: National Academy of Sciences.

DOCUMENT CONTROL DATA

(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document, Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's document, or tasking agency, are entered in section 8.)

Publishing: DRDC
Toronto

Performing: Humansystems® Incorporated, 111 Farquhar St., 2nd floor, Guelph, ON N1H 3N4

Monitoring:

Contracting: DRDC
Toronto

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification is indicated by the appropriate abbreviation (S, C, R, or U) in parenthesis at the end of the title)

Voice Controls and Displays for the Dismounted Soldier (U)
Systèmes de Commande et d'affichage Vocaux Destinés aux Fantassins Débarqués

4. AUTHORS (First name, middle initial and last name. If military, show rank, e.g. Maj. John E. Doe.)

L. Thompson; David W. Tack

5. DATE OF PUBLICATION (Month and year of publication of document.)

October 2005

6a NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

37

6b. NO. OF REFS (Total cited in document.)

46

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of document, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (The names of the department project office or laboratory sponsoring the research and development − include address.)

Sponsoring: DLR 5, NDHQ OTTAWA,ON K1A 0K2

Tasking:

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant under which the document was written. Please specify whether project or grant.)

12QG01

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

W7711−017747/001/TOR

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document)

DRDC Toronto CR 2005−033

10b. OTHER DOCUMENT NO(s). (Any other numbers under which may be assigned this document either by the originator or by the sponsor.)

SIREQ #90

11. DOCUMENT AVAILABILIY (Any limitations on the dissemination of the document, other than those imposed by security classification.)

Defence departments in approved countries − Document has initial limited distribution through Exploitation Manager − TTCP and NATO countries and agencies − Unlimited after initial limited distribution

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11), However, when further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Other − Document to have initial Limited announcement

DOCUMENT CONTROL DATA

(Security classification of the title, body of abstract and indexing annotation must be entered when the overall document is classified)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

(U) The purpose of this report is to investigate the potential application of voice recognition and speech synthesis to a wearable computer for use by dismounted soldiers. A literature review was conducted to identify the issues and capabilities of voice controls and displays for dismounted infantry soldiers. Current applications of voice interfaces are identified including the Force XXI modified Land Warrior user voice control system. Considerations regarding the appropriate use of speech input and voice displays are discussed including use of speech as opposed to tonal or visual displays, text−to−speech synthesis, types of speech displays and speech display intelligibility. Factors influencing the effectiveness of speech input and voice control systems are described including speaker dependence, word boundaries, vocabulary size, and system performance. Environmental factors affecting speech recognition accuracy are discussed including ambient noise, vibration, and operator stress. Finally, this report considers general system design issues including speech input device activation and deactivation, multi−modal displays and controls, feedback and error correction, vocabulary size and selection, robustness to ambient noise, system size, microphone and headset design, compatibility with masks and exclusion garments, training, and security issues.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

(U) Soldier Information Requirements Technology Demonstration Project; SIREQ TD; voice recognition; text−to−speech; speech synthesis; speech input; voice controls