Force XXI Land Warrior: A Systems Approach to Speech Recognition

C. C. Broun and W. M. Campbell
Motorola Human Interface Laboratory
Tempe, AZ 85284, USA

ABSTRACT

Speech recognition is continually being realized as a user interface in new applications. As this technology progresses, it enables new ways for humans to interact with machines and information. The performance in many domains has approached users' expectations. Although there are still abundant technology challenges ahead, speech recognition has reached a maturity level that requires one to consider its deployment in complex systems and environments. It is in this vein that we discuss a systems approach to the successful execution of speech recognition within the Force XXI Land Warrior program.

We discuss the System Voice Control component as it fits within the overall program. The requirements for robustness, recognition, and computational complexity issues are addressed. We explicitly cover the system aspects and how they influence the user interface and reveal the parameters for actual use. Finally, we consider the implementation of a polynomial-based classifier for speech recognition, and we provide the final system performance measures on a large domain specific database.

1. INTRODUCTION

The Land Warrior Engineering Manufacturing Development (EMD) program is the Army's revolutionary program to develop and field a totally integrated Soldier Fighting System. This system uses advanced technologies to render unparalleled effectiveness by providing an improved capability to detect, acquire, locate and engage targets at greater ranges, day or night. The system links the individual soldier to the digitized battlefield for improved communications and situational awareness.

The purpose of the Force XXI Land Warrior program is to accelerate the fielding of advanced technology upgrades to the Land Warrior EMD platform. This ensures a global technology advantage for dismounted warrior combat systems.

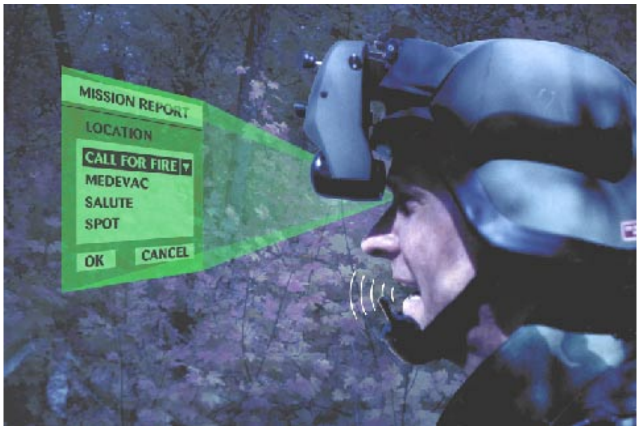

Figure 1: System Voice Control concept.

The System Voice Control (SVC) component of Force XXI Land Warrior provides a speech interface to the existing soldier computer from the Land Warrior EMD program. The intent is to provide the dismounted soldier with an efficient method of *hands-busy*, *eyes-busy* control of the soldier system. Figure 1 illustrates the SVC concept.

The application of speech recognition in a combat environment elicits challenging performance requirements. Both recognition and out-of-vocabulary (OOV) rejection must maintain usable performance levels in adverse noise conditions. In addition, the system must respond to a wide dynamic range of voice levels – low levels for covert operations, and high levels for noisy situations. Voice stress is another concern for performance, as the Lombard effect [1] is often encountered. Finally, the algorithms must be computationally efficient so as not to drain the system battery, and the word models must be sufficiently small in order to fit into the available memory.

The design of the overall system and the user interface is discussed in Section 2. The structure of the actual fielded speech recognition algorithm, which is based on a polynomial classifier, is explained in Section 3. In Sections 4 and 5, solutions to the technical problems of noise robustness, stressed speech and out-of-vocabulary rejection are discussed. Finally, the validation and results of the final system are given in Section 6.

2. SYSTEM DESIGN & USER INTERFACE

The basic block diagram of the SVC component is shown in Figure 2. The soldier depresses a button on his weapon to initiate the recognition system. A close-talking noise-canceling microphone captures the spoken command, which is digitized by an A/D. The recognizer processes the sampled speech, and the appropriate response is generated by the soldier computer.

The successful deployment of speech recognition is credited with the systems approach to the design and implementation of SVC. Each component, from the user interface to the back-end classifier, is designed to work together in order to maximize performance for this application and for the target users.

Understanding of the users' environment and true use scenarios is gained through extensive user experience studies and actual involvement in live-fire (and other) exercises. Based on the collected data, one extracts detailed user interface specifications as well as the expected environmental conditions in which the speech recognition algorithms must operate.

One of the more significant user interface issues focuses on the method of initiation of the speech recognition engine. There are two criteria. The soldier must be unencumbered, and the system

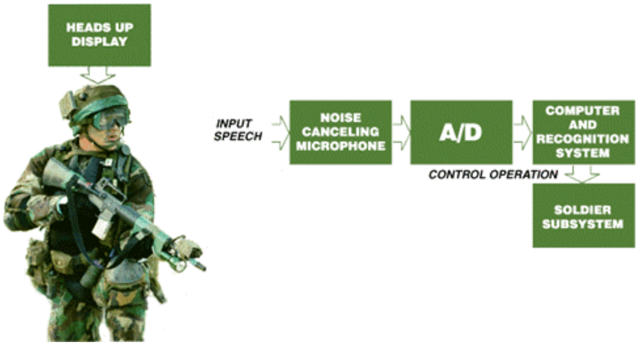

Figure 2: System Voice Control system block diagram.

accuracy must meet his expectations. A keyword-based system would allow completely hands-free operation, but has two distinct drawbacks. First, the keyword algorithm must run constantly, which introduces a drain on the battery. Secondly, such a system is susceptible to insertion errors, or false command recognition. In order to resolve these issues, a push-to-talk interface is favored.

For the push-to-talk interface, the location of the button is critical to the soldiers' ability to easily initiate the recognizer. Several options were proposed, including the addition of a button on an advanced weapon. With this combination of button placement and the use of speech recognition as the primary interface, it is clearly demonstrated that the soldier can increase his effectiveness. Figure 3 contrasts a soldier using the speech interface (left) with a soldier using a traditional pointing device (right). The soldier using SVC keeps both hands on his weapon and maintains focus downrange, whereas the other has his weapon pointed into the ground while manipulating the mouse.

In addition to the importance of the user interface, constraints on power consumption and robustness affect the choice of fielded speech recognition algorithms (Section 3). In order to provide the required performance, the software is tightly integrated with the overall system with specific design criteria to handle adverse noisy environments and stressed speech (Section 4), as well as noise robust OOV rejection (Section 5).

3. CLASSIFIER STRUCTURE

Many classification methods are currently being applied to the problem of speech recognition. Traditionally, statistical methods are used to model the speaker's speech data; the most popular approach is the Hidden Markov Model (HMM). More recently, discriminative classification techniques have been applied to the problem. In order to provide the best performance for speech recognition systems, they include out-of-class data in the training phase. For SVC, an approach based on polynomial classifiers is implemented.

The basic structure of our classifier is shown in Figure 4. The feature vectors, $\mathbf{x}_1 \ldots \mathbf{x}_M$, are input into a discriminant function, $\mathbf{w}^t\mathbf{p}(\mathbf{x})$, and then the output is averaged over all M to produce a score. This strategy is similar to hybrid connectionist approaches where artificial neural networks are used as observation probability generators [2].

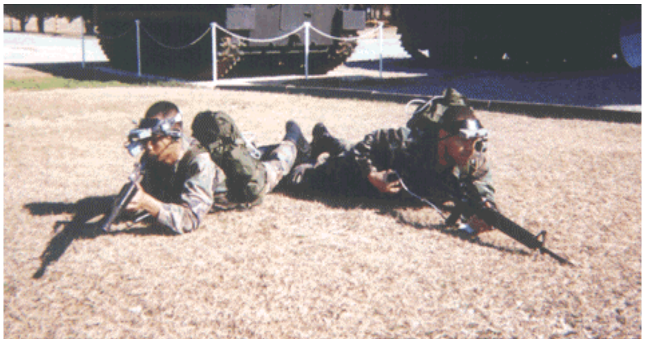

Figure 3: User interface testing.

Our pattern classifier uses a polynomial discriminant function

$$f(\mathbf{x},\mathbf{w}) = \mathbf{w}^t \mathbf{p}(\mathbf{x}) . \tag{1}$$

The discriminant function is composed of two parts. The first part, $\mathbf{w}$, is the model for the particular class. The second part, $\mathbf{p}(\mathbf{x})$, is a polynomial basis vector constructed from input feature vector $\mathbf{x}$. This basis vector is the monomial terms up to degree K of the input features. Thus, the discriminant function output is a linear combination of the polynomial basis elements.

For speech recognition, it is important for scoring to be computationally efficient. The soldier system load is determined by the complexity of the discriminant function evaluation. Since $\mathbf{w}$ does not depend on the frame index, scoring can be simplified as follows:

$$s = \mathbf{w}^t \frac{1}{M} \sum_{k=1}^{M} \mathbf{p}(\mathbf{x}_k) = \mathbf{w}^t \overline{\mathbf{p}} . \tag{2}$$

Only a single vector representing the input speech is computed, and each score evaluation equates to computing an inner product. The number of floating point operations (FLOPS) is $2N_{\text{model}}$-1, where N_{model} is the length of $\mathbf{w}$.

Thus, for 15 features and a 3rd order (K = 3) polynomial expansion, $\mathbf{w}$ is of length 816, resulting in only 1631 FLOPS per word score, and a model size of 3264 bytes for a floating point representation. A detailed description is given in [3].

4. NOISE & STRESSED SPEECH

To ensure that SVC performs reliably for the dismounted soldier, domain specific noise data from live fire exercises is evaluated. The significant noise sources are vehicle and weapons fire. Analysis of this data (and similar noise data from the NOISEX-92 database) provides an understanding of the specific spectral characteristics that are expected in real use scenarios. The typical

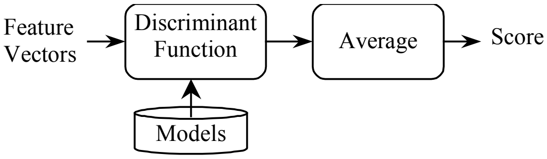

Figure 4: Classifier structure.

Table 1: Stressed speech recognition performance for the SUSAS database. (Lom = Lombard; Ques = Question)

Condition	Stressed Speech Domain (Speaker Independent)									
	Slow	Fast	Soft	Loud	Angry	Clear	Ques	C50	C70	Lom
No Compensation	77%	79%	77%	61%	56%	80%	78%	81%	78%	69%
CMS & Var. Norm.	81%	81%	83%	74%	70%	86%	82%	87%	87%	81%

vehicle noise is characterized as quasi-stationary with a low-pass spectral envelope. The weapons fire is primarily impulsive.

Several acoustic tests are performed with several microphones in an effort to characterize the overall system response. A flat response transducer is used as a baseline, and two noise canceling microphones are tested. The first is an Andrea, with which we have experience; the other is a Gentex electret element, which is the communications microphone used in the Land Warrior EMD system.

Examples of vehicle noise and machinegun fire are played through loudspeakers and recorded, both by itself and with an individual speaking commands. Analysis of the recorded utterances provides the spectral characteristics of the near-field and far-field responses of the microphones. From this data, an interpolated finite impulse response (IFIR) filter is designed to suppress the low frequencies dominant in the noise sources.

The primary concern for stressed speech is the Lombard effect [1]. The dominant characteristic is a tilt in the spectrum, similar to channel mismatch. To compensate for this anomaly, cepstral mean subtraction (CMS) and variance normalization are employed. The results on the SUSAS database [4] are shown in Table 1.

5. OUT-OF-VOCABULARY REJECTION

The approach for OOV rejection must perform well in noisy conditions without significantly degrading the recognition performance. In addition, memory, computational complexity and latency issues are of concern.

The out-of-class rejection algorithm is given in Table 2. The idea is to populate the feature space with a large number of out-of-class (garbage) models. This tends to tighten the distributions around the in-class models, and identifies out-of-class regions between the in-class regions. This is similar to the ambiguity rejection criteria, but does not rely upon distribution based thresholds.

Table 2: Out-of-vocabulary rejection algorithm.

1. Generate scores for all in-class and garbage models
2. Sort the scores such that the highest score is first (rank=1)
3. Find the highest scoring in-class model, **w***
4. Eliminate the cohorts for this model (**w***) from the sorted score list
5. Get the overall rank, *r*, for this model (**w***) in the sorted score list
6. If this rank, *r*, is less than or equal to the rank tolerance, *R*, then output **w*** as the recognized class; otherwise reject the input

In order to allow for the variation of input features in mismatched conditions, a rank tolerance is used. This rank tolerance allows for varying statistics of the in-class feature vectors without generating false rejects.

For SVC, this rejection criterion requires no additional memory to store garbage models, and only minimal computational cost to generate the garbage scores. The application is broken into different subsets (menus). For each subset, the *non-active* vocabulary models are used as garbage classes. This structure averts the need for additional memory for distinct garbage models. Since the scoring of the discriminant function is simply an inner product, the added computational cost of scoring all models for each input utterance is insignificant compared to the initial feature extraction.

This technique is extensively tested, and the performance is shown to be substantially more robust than standard threshold methods. For a detailed description of the algorithm and its performance characteristics see [5].

6. VALIDATION

We perform our experiments on a domain specific isolated word database. It consists of short command phrases that are used to control the soldier computer. In total there are 131 unique in-vocabulary phrases that are split into 27 subsets (contexts) ranging in perplexity from 2 to 35. The database is collected at a sampling rate of 11.025 kHz and 16-bit mono using the Gentex microphone. The training set consists of 100 male soldiers repeating each command once. The test set consists of 66 male soldiers repeating the same commands one time. The speakers in the training and testing sets are distinct, and they have a wide variety of accents.

The experiments are performed for four environments. The *quiet* environment is the matched condition (i.e., the speech is not modified in any way from the original recording). The other three environments represent typical noisy conditions. They are constructed by mixing noise from the NOISEX-92 database with the test speech utterances only; the training speech utterances remain unaltered. Noise from the *leopard*, *m109*, and *machinegun* sources are used. The signal-to-noise ratios of the test phrases are given in Table 3.

Table 3: Average SNRs for test phrases.

Source	Average SNR
Quiet	33 dB
Leopard	14 dB
M109	13 dB
Machinegun	28 dB

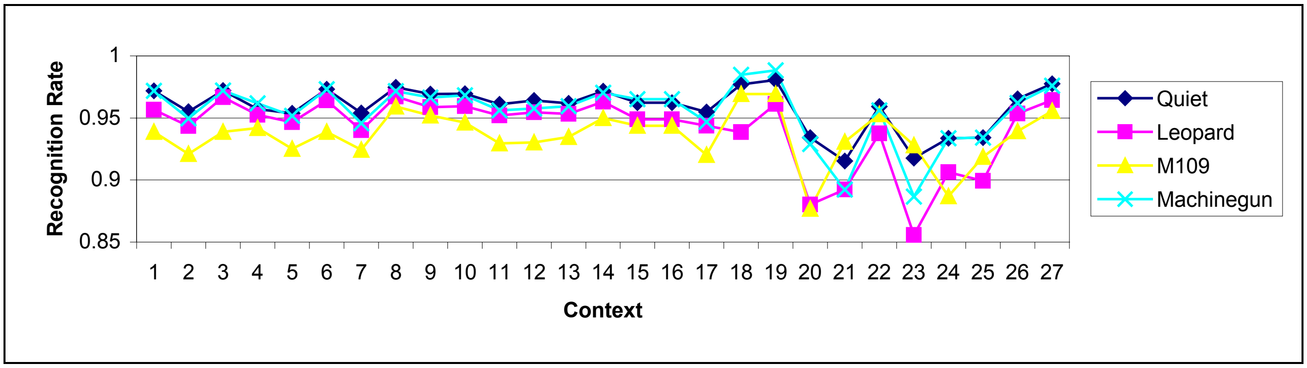

Figure 5: Speech recognition performance for all contexts.

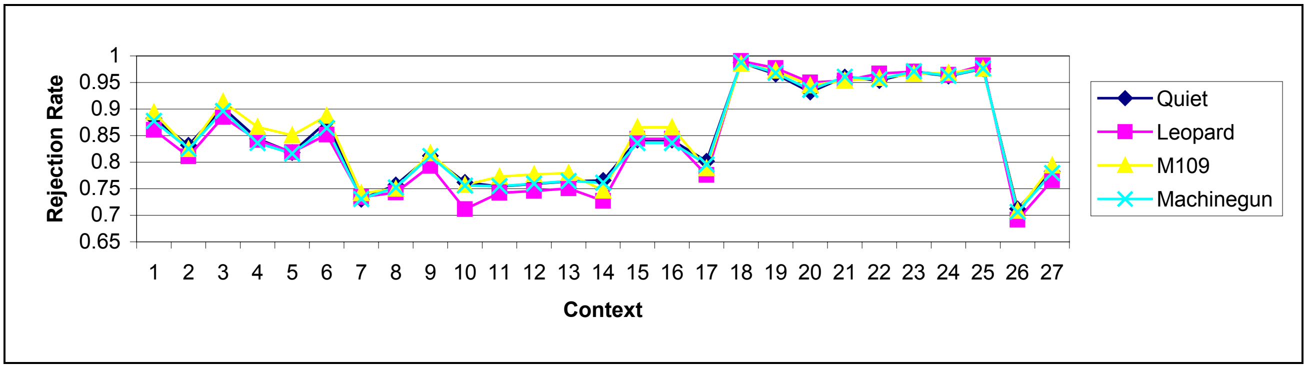

Figure 6: Out-of-vocabulary rejection performance for all contexts.

A 3rd order polynomial system is implemented. We analyze the speech using 30 ms frames. Mean removal followed by a IFIR filter and a Hamming window is performed at 100 frames per second. We use 12th order LP analysis and then derive 14 cepstral parameters. Endpointing is performed using frame energy; feature vectors corresponding to non-speech frames are discarded. Cepstral mean removal and variance normalization are applied to the features to compensate for stressed speech. Additionally, a normalized-time feature, i/N_{frames} is appended, for a total of 15 features. The resulting feature vectors are then input to the system.

Performance for each of the 27 contexts is evaluated. The recognition accuracy (with OOV rejection enabled) is shown in Figure 5. Of interest is the fact that the performance does not significantly degrade in the presence noise.

The out-of-vocabulary database consists of 118 phrases (distinct from the in-vocabulary phrases) and 3 spurious inputs (breath, lip-smack, and cough). All of the speech is collected in a manner similar to that described above.

Again, the performance for each of the 27 contexts is evaluated, (Figure 6), and we see that the OOV rejection is robust to noisy conditions.

7. CONCLUSIONS

Successful implementation of speech recognition interfaces is achieved via a systems approach. It is paramount that users are brought into the design process as early as possible in order to fully capture their expectations and the system use parameters. It should be considered standard practice to customize solutions for new domains and environments.

8. REFERENCES

[1] J. C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” in *J. Acoust. Soc. Am.*, vol. 1, pp. 510-524, 1993.

[2] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[3] W. M. Campbell, K. T. Assaleh, and C. C. Broun, “Low-Complexity Small-Vocabulary Speech Recognition for Portable Devices,” in *Proceedings of the Fifth International Conference on Signal Processing and its Applications*, pp. 619-622, 1999.

[4] J. H. L. Hansen and S. E. Bou-Ghazale, “Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database,” in *Proceedings of Eurospeech*, vol. 4, pp. 1743-1746, 1997.

[5] C. C. Broun and W. M. Campbell, “Robust Out-of-Vocabulary Rejection for Low-Complexity Speaker Independent Speech Recognition,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1811-1814, 2000.