CIVILIAN RESEARCH PROJECT

Advancing Noise Robust Automatic Speech Recognition for Command and Control Applications

By

Colonel JAMES D. BASS, Ph.D.
United States Army

Mr. Robert Riffle
Coordinating Advisor
The University of Texas at Austin

US Army War College
CARLISLE BARRACKS, PENNSYLVANIA 17013

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
28 MAR 2006	**Civilian Research Project**	**00-00-2005 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Advancing Noise Robust Automatic Speech Recognition for Command and Control Applications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
James Bass	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army War College,Carlisle Barracks,Carlisle,PA,17013-5050	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Institute of Advanced Technology, Att: Robert Riffle, University of Texas at Austin, 3925 West Braker Lane, Ste 400, Austin, TX, 78759-5316	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
See attached.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**29**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

ABSTRACT

AUTHOR: Colonel James D. Bass, Ph.D.

TITLE: Advancing Noise Robust Automatic Speech Recognition for Command and Control Applications

FORMAT: Civilian Research Project

DATE: 31 March 2006 PAGES: 23 CLASSIFICATION: Unclassified

This is a technical assessment paper intended for use by engineers and research scientist working on the development and integration of Automatic Speech Recognition (ASR), it will cover the state of speech and recognition technologies with emphasis on noise robust command and control (C^2) application. The reliable elimination of the keyboard and mouse in mounted and un-mounted C^2 systems has been a desire of systems developers and requirements writers since the development of PC-based ASR systems in the early 1990's. However, current research and commercial quality ASR applications never had the noise robustness to support a truly tactical C^2 application. As ASR achieved limited operational success in noisy environments around the 2002 timeframe, the C^2 requirements evolved to include the emerging system of systems approach and multilingual operational environments in support of the Global War On Terrorism (GWOT)—in such environments, the system must understand not just words as commands (ASR), but to understand phrases and sentences (semantic and syntactic) and reply in a conversational manner (speech and natural language generation). If the keyboard and mouse are to be truly eliminated, a system now needs to conduct a natural conversation with an operator and possibly others in the operational environment. This paper will cover the advances, limitations, and reasonable expectations from several levels: Research Scientist and Engineers, Program Executive Office (PEO), Program Manager (PM), and requirements office. I will also discuss the major technical challenges that remain as well as some risk assessment to help decision makers align expectations with reasonable availability dates based on current and future research efforts. Our user and requirements communities have waited 20 years for these technologies to mature to a level of effectiveness. Now it is time to review and assess the available research and commercial products to see what is available for use in the tactical environment.

Table of Contents

ACKNOWLEDGMENTS

This paper is the result of the author's participation at the U.S. Army War College Fellowship program at the Institute for Advanced Technology at The University of Texas at Austin. I wish to express my thanks and gratitude to the faculty, researchers, and staff of this fine research institution.

1.0 Introduction

1.1 The promise of ASR

Command and Control (C^2) on the move, is a goal that has been discussed and desired since the introduction of communications devices on vehicles. With the consistent advancement in computer power (following the timelines of Moore's Law [1]), new innovations supporting robust Automatic Speech Recognition (ASR) and advances in Digital Signal Processing (DSP), this goal is evolving into a reality. As the technology advanced, it has faced many difficult obstacles. Human speech recognition is a science still not fully understood by both the science and medical communities. How can humans filter multiple voices and "concentrate" on a single voice—even when it is at a much lower sound level than other competing voices and sounds? How does the cochlear component of the human ear play in this process? Researchers cannot even decide on the fundamental starting question, "Should computers be taught to hear like humans or use a purely artificial digital—based paradigm?" However, despite these difficult problems, they pale in comparison to the most fundamental problem faced today by the ASR research community—botched management of expectations. Based on the promises of the DSP revolution of the 1980's, expectations in ASR research soared. Funding from Defense Advanced Research Projects Agency (DARPA), National Science Foundation (NSF), and even venture capital allowed centers of excellence to be established at many research universities and prominent commercial research institutions. The race was on to not only research, but develop practical ASR products for use by the military, industry, and the physically challenged.

1.2 The seeds of disappointment

The initial ASR research was conducted on super computers in non-real-time with little expectation for small platform real-time ASR application. At this stage it was a purely academic adventure in basic research. However based on the phenomenal efforts of the Very Large System Integration (VLSI) community, the processing power of the small platform computer exploded during the 1990's. Moore's law predicted a doubling of computational power for a given platform approximately every eighteen months. His prediction held true throughout the 90's and is expected to hold for another ten years [2]. Research in massively parallel processing, Quantum Computing, and Bio Computing, all promise to keep the advances in processing power

moving forward. During this same time frame, the military requirements community called for the development of real-time or near real-time ASR products to support basic on-the-move C2 operations in garrison and the field. In addition to the military demands, commercial companies like IBM and Dragon anticipated a large market for ASR products. The results, based on these expectations, were massive developmental expenditures in the 1990's designed to deliver PC-based ASR products.

The results of both the military and commercial research and development programs were a profound disappointment. At the research level, the new algorithms and DSP techniques showed a steady linear improvement; however, in attempts to transition the technology to users, the software was unstable, prone to crashing, demanded almost exclusive use of the CPU, was difficult to integrate into user applications (API's were also unstable), and the latency was too great for most real-time application. In addition to these issues, the accuracy of the ASR engine was only in the 90% - 95% range (very poor for practical use) and dropped to unacceptable levels with the introduction of any form of noise (stationary or non-stationary) even with sophisticated noise canceling technologies. Introduction of these software products was disastrous. With user expectations shattered and promised performance levels unmet, the ASR R&D community experienced a massive drop in research and developmental funding. The commercial community saw venture capital dry up along with a collapse of commercial sales. By 1999 there was only one major DARPA ASR program left, *Communicator*, and several ASR commercial developers were forced to merge for survival and while others completely closed shop.

1.3 A call for reassessment

Out of the ashes of these unsuccessful efforts, a call for an assessment was made by the leading researchers in the various fields that supported ASR. Their goal was to establish the next generation of research and to repair the reputation of the community. This group of researchers and engineers held this meeting in December 1999 under the auspices of the IEEE ASRU Workshop (the IEEE Signal Processing Society has sponsored ASRU biannual workshops on various topics since 1989). It was at this meeting that a complete reassessment was openly discussed and new research paradigms were debated.

1.4 Roadmap for achieving the promise of ASR

At the time of the ASRU workshop in 1999, the major research in ASR was concentrating on use of Hidden Markov Models (HMM), adaptive systems for speaker independence (systems that do not require user to read sample speech to calibrate the system), large vocabularies, use of far field microphones, systems resilient to multi-speakers interference, and use of ASR in conversational telephony based systems. The community had become caught up in a game of metrics showing only modest improvements. The term *"local optima"* was the refrain heard throughout the ASRU meeting [3]. Subsystems were being optimized for small improvements in performance, but the overall system showed barely statistically significant improvement [4]. At ASRU the old template based and current statistical approaches were challenged and new feature-based methodologies were introduced. The result was an invigorated community given a chance to develop new approaches through several new funded research programs: the DARPA *Communicator*, *TIDES*, and *Babylon* programs [5].

2.0 State of the Technology

2.1 Review of State of the Practice Systems

ASR is a complex system of software processes, hardware devices and computational engines. From the input microphone and digital processing of those utterances to the generation of sound or text (as required by the application), each step must be integrated emphasizing accuracy and speed. In state-of-the-practice commercial and current militarized systems, the processes are well defined. Major differences in commercial vs. military are variations in corpus and vocabulary and the integration of military input devices and speech filtering techniques such as those used in military vehicle intercoms. Here is a brief outline showing the major processes of ASR as a complete system. Proprietary solutions exist between vendors and developers, but most of these "state of the practice" systems deliver statistically similar results with similar fundamental methods [6]:

1. Speech Model
    a. Mapping of speech production
        i. Vocal tract
        ii. Voiced and unvoiced speech

iii. Articulatory phonetics
iv. Phoneme development
v. Prosodic modeling
vi. Dialect compensation

b. Representing speech in a computer
i. Input devices (microphone, inducer, etc.)
ii. Audio sampling
iii. Quantization
iv. Speech digitization
v. Wave coders
vi. Voice coders
vii. Transforms (DFT, FFT, etc.)

2. ASR Engine
a. Mapping system performance parameters
i. Continuous vs. non-continuous speech
ii. Speaker independence vs. speaker dependence
iii. Vocabulary size
b. Acoustic feature selection
c. Comparing features
i. Dynamic Time Warping
ii. Hidden Markov Models
iii. Other methods
d. Error detection and correction
e. Learning system techniques

3. Language Model
a. Corpus collection or generation
b. Bi, Tri, and N-gram techniques
c. Confusion matrix

4. Speech/Text Generation
(NOTE: Not all systems generate speech, some are text only.)
a. Method selection
i. Parametric
ii. Concatenative
b. Text-to-Speech processing
i. Rule and exception processing
ii. Morphological analysis
iii. Articulation effects
iv. Prosody

In a relatively quite environment such as an office setting (=13db Signal to Noise Ratio) with a close talking microphone, users of theses systems with speaker training (speaker dependent system), with a vocabulary in the 80K word range can expect approximately a 90% accuracy rate [7]. Unfortunately that error rate is usually not tolerated by normal officer workers. Follow-on system training and user corrections added to the learning system usually boost performance by

only 2-3% [8]. In industrial and military applications where the noise level can easily exceed 30db, system performance quickly degrades to less that 50% word accuracy. Training in the noise yields some increase in performance; however, in the noise has some variation, the performance increase is not reliable. In addition to accuracy, the latency of receiving the results of these systems is unacceptable for real-time or near-real time applications. Thus the standard commercial systems have minor application in sterile office environments or use by the physically challenged. To address the shortcomings of the commercial quality systems, DARPA, Naval Research Lab (NRL), Army Research Lab (ARL), and the NSF all funded basic research in new approaches to ASR in noisy environments.

2.2 Review of State of the Art Systems

State of the art noise robust ASR systems are being designed for both specific noise environments and as general adverse condition systems. The key research domain for these systems is adaptive methodologies. Systems are created with samples of noise that are expected in the operating environment (e.g., automobiles, aircraft, military vehicles, etc) [9]. The speech models are enhanced by incorporating prosody (speech variation in pitch and speed usually based on stress) and attempt to evaluate the users stress level to adapt the feature interpretation to support higher accuracy rates. Other system developers are concentrating on adaptation at the input level. These developers are looking at technologies for noise cancellation, steerable microphone arrays, use of lip reading technology, and other novel approaches [10].

Beyond adaptation a few researchers are looking at enhanced speech features for enhancing the accuracy and noise robustness of ASR systems [11]. These researchers are using such methods as Maximum A Posteriori (MAP) or the use of clustering techniques with spectral vectors of clean speech [12]. These techniques allow noise corrupted components of noisy speech to be identified and cleared. These noise robust systems developed by such organizations as BBN, IBM, Carnegie-Mellon, SRI, and MIT all show superior noise robust capabilities over commercial systems; however, their latency, restricted vocabularies, and/or computational and hardware requirements make practical application of these systems in the operational environment difficult and expensive.

2.3 Assessment

For commercial application current state of practice systems have found relative success in the automated conversational telephone system market. Spurred by the DARPA *Communicator* program, there are several hundred of these conversational systems operating in Europe and the United Sates [13]. These systems take advantage of the fact that the user is calling to complete a finite set of tasks. Thus the vocabulary can be constrained to some extent by the domain. As the user gets deeper into task completion, the domain becomes more constrained—thus enhancing its accuracy and speed as it collects information from and provides information to the user.

For military application, the only reasonably effective applications of ASR in a (non-office) military environment have been in limited domain devices such as the *Phraselator* [14]. This small hand-held translation device uses a non-continuous speech (set phrases), speaker independent, small vocabulary paradigm. The ASR is used to identify one of several thousand fixed phrases that are tied to set of identical fixed phrases in a foreign language. The user speaks an English phrase and the system transmits the same phrase in one of several languages (e.g., Arabic, Pashto, or Urdu). This PDA sized system has been procured and used with some success in the harsh environments of Afghanistan and Iraq. For a Command and Control (C^2) application in the harsh noise of an operational environment, the current crop of commercial and developmental systems could only be effective in limited domain directive (command based) applications.

3.0 Hard Unsolved Problem Domains

3.1 Acoustic Domain

Noise is fundamentally defined as, "Sound, or a sound that is loud, disagreeable, or unwanted" [15]. The major acoustic objective in ASR involves the detection and enhancement of speech features and the degradation or elimination of noise. There are three sources of noise in speech recognition: the speaker, the background, and the channel(s) [16]. Within the ASR acoustic domain, noise from those sources fall into three categories: channel noise, ambient noise, and processing noise. In commercial ASR systems, the minimal acceptable Signal-to-Noise Ration (SNR) is approximately 10db. With a conventional cardioic microphone, a well trained operator and system can achieve 85-90% accuracy rates under ideal conditions. This

SNR with microphones and standard audio filters is achievable in relatively quite commercial office environments, but are very difficult to achieve in any operational military environment. The current technology thrusts have focused on these three major categories of noise. As is the nature of potential solutions, each alternative brings resource requirements that must be accommodated in the target system.

3.1.1 Ambient Noise

Ambient noise can come from the external environment or even the operator of an ASR system. Ambient noise can be steady (periodic) or constantly varying (non-periodic). Both forms (generally labeled as stationary and non-stationary noise) must be significantly degraded or eliminated for current technology ASR systems to work. There are four general approaches to attack this category of noise: noise canceling technologies, highly directional microphones, steerable microphone arrays, and alternative filtering (most noticeable of late cochlear filters) [17]. Each of these technologies offers help in raising the effective SNR by either notching out noise, enhancing desired speech components (features), or a combination of the two.

All of these potential solutions are limited by their impact on the operator, the system, or external restrictions. For example highly direction microphones are only effective if the user is able to stay within the very narrow beam of the microphone. Steerable arrays are good on vehicles or in conference room environments. Noise cancellation technologies are advancing well, but for real-time systems noise canceling systems introduce small bits of noise created by the delay in sampling the sound and then creating the canceling signal. These small noise features are referred to as artifacts and take on a random appearance. Such artifacts fall within the category of processing noise and must be dealt with by the ASR system. These artifacts, though a nuisance, pale in comparison to the general system performance increase provided by noise canceling technologies—especially in stationary noise. The last proposed solution is the most immature and least understood. Filtering is fundamentally the feeding of all frequencies received by the input device, but outputting only certain frequencies. Basic filter types include high-pass, low-pass, and band-pass filters.

Modern filters are far more flexible and are able to mimic the type of filtering done in biological systems—such as the human ear. Though not fully understood, the cochlear filter in humans can be emulated to some degree with modern non-linear digital signal processing technologies. Early linear types offered only marginal improvements in speech feature

enhancement; however, with the introduction of non-linear processing, the characteristics of the artificial cochlear is much closer to that of its biological counter part. The primary issues with this solution are system latency, enhanced processing, and power requirements. At this time the maturity of Cochlear filters offers the greatest potential for enhancing speech in ambient noise. Research now revolves around the mix and integration of all of these technologies and the impact on system design, power, latency, and thus fundamental usability. For the foreseeable future, any practical implementation of ASR in an operational environment will use one, a combination or all of these techniques.

3.1.2 Channel Noise

The source of most channel noise is typically introduced by the microphone or transmission medium and manifests itself as hum or static. Sound engineering practices to eliminate ground loops and good shielding of external signals usually eliminate these problems. The RF rich environments of the military usually require much more attention than typical commercial office environments—though with the introduction of WiFi, Bluetooth, and other wireless technologies, even commercial applications need a good understanding of their RF environment. Another source of noise is variation caused by a mismatch between channels. The most common channel mismatches are introduced by variations in band-pass filters or reactive resistance to the physical transmission lines (commonly measured as the Standing Wave Ration (SWR) [18]. Again sound engineering practices and training operators to use the correct cables and microphones usually eliminate this form of channel noise. Of all the categories of noise, these should be the least tolerated by the community. Channel noise is a reflection of good engineering practices and knowledge of the operational environment.

3.1.3 Processing Noise

The most critical form of processing noise in ASR is introduced through quantization of the speech signal for digital processing. Noise in this context is defined as unwanted deviations from the original signal. Quantization noise may be represented in decibels as a signal-to-noise ratio [19]. SNR is measured as the relative cumulative errors in a reconstructed signal and can be improved by increasing the quantize—at the cost of increased resources (e.g., power, computational support, etc.).

The sampling rate and the quantizer size are the two primary parameters of sampling; these two factors determine how faithfully the original signal can be recovered after the analog-to-digital and digital-to-analog process. These two parameters are independent of each other with respect to fidelity of the digitized signal. An increase in sampling rate will not reduce the overall quantitization error, and aliasing will occur if the sampling rate is too low—regardless of the size of the quantizer. The development of new nonlinear quantizers such as logarithmic and adaptive seek to reduce overall quantization error by concentrating the greater part of their quantization levels in information rich audio ranges in the hopes of achieving better resolution at the expense of poor resolution in those other audio ranges—an example would be a non-linear quantizer optimized for the human audio range. The elimination of processing noise is very much an art at this stage of development in ASR. The identification of new or alternative speech features is usually associated with a modification of the quantizer to extract the information within that feature.

As the multidisciplinary research community tries to identify the most critical speech features for exploitation, the digitization community must continue its development of new quantization routines to keep processing noise at the absolute minimum. So far the history of their efforts aiding the ASR community has proven effective.

3.2 Speech Extraction

We have a signal; the signal has been digitally filtered to remove as much noise as possible while preserving as much of the speech component as practical. What's next? All computer speech recognition systems are based on training the computer system by exposing it to known speech from which a reference model is created. Then the new unknown speech is compared to the reference model to produce a hypothesis solution (translation) which is then validated by various support modules further along of the ASR process. The reference models may be built on speech units as large as a phrase or as small as a fraction of an allophone. There are tradeoffs based on these choices. After a speech model is selected, a unit of representation must be selected to allow robust comparison of new speech against the reference model, here again a brief discussion of major methods and their advantages and disadvantages need to be covered. The final major component of the speech extraction process is the method by which the

units will be compared. There are currently two major methods with a new novel approach on the horizon, all three will be covered.

3.2.1 Words and Phrases

Early recognizers used single words or groups of words spoken together (phrases) as the basic unit of speech. Developers quickly recognized that some common word pairs or phrases when built together in the reference model improved overall system accuracy. For example a phrase such as *come here* would be entered as a single unit instead of the two separate units of *come* and *here*. The reason is that the more dynamic components a unit of speech has, the better chance of a successfully recognition. In such utterances as *the cook*, *the wife*, *the thief*, etc. the *the* in these examples are usually poorly articulated and would be missed by most recognizers. Designers learned to create such word pairs and phrases to add to the reference model—thus increasing the accuracy of the model at the cost of model size and reference model processing time.

In the early days of ASR development, such brute force techniques as word and word phrase modeling were practical only for very small vocabulary applications. Storage space for the reference model and computational horsepower for access and comparison were not powerful enough to overcome the system latency of these prototypes. They were simply too slow and too limited for any practical application.

3.2.2 Syllables

By building the reference model from components of words instead of words and phrases, designers can reduce the size of the model. Words are composed of syllables and a great amount of redundancy can be reduced. This efficiency is very language dependent. Languages like English with more than 10,000 syllables realize a modest efficiency; however, Japanese and other languages built on only a few hundred syllables can experience a substantial reduction in the reference model size. Syllable based reference models have fallen out of favor do to the fact that most speakers in all languages alter the classic pronunciation of suffixes and predicates based on long term social or cultural norms. For example, the suffix *–less* (e.g., *godless*) tend to be pronounced with a schwa E (?). This is but one example of thousands, and

they exist in all languages. This results in the creation of thousands of exceptions to the basic syllable rules. Worse yet, they tend to be applied by speakers on a near random basis. Even low syllable count languages like Japanese are subject to these same events. Unfortunately, when all of the variations are accounted for, the hoped for efficiency in reference model size becomes quite modest, leading researchers to look for alternative solutions.

3.2.3 Phonemes

As indicated by some researchers and authors, the term phoneme-based recognition is actually more correctly described by the label allophone-based recognition. Since an allophone is the smallest unit of sound that is actually produced for speech. Similar to the hopes in the development of syllable-based models, the hopes in phoneme-based models was that by using the basic sound elements of speech, a highly efficient reference model could be designed. Most languages in the world are based upon approximately 100 allophones.

Unfortunately, the interaction problems (e.g., interallophone articulatory effects) as experienced in the syllable-based model, as well as others issues not anticipated, have forced researchers to look beyond the phoneme-based model as a potential solution. The major problem involved in the use of phoneme-based models is the fundamentally infinite variations in the actual sound of each phoneme. Also, the interactions with adjacent phonemes create unique sounds – worse yet, these sounds are not consistent. The problem of infinite variation was overcome to some extent by the use of cardinal points of identification. An example would be the presence or absence of aspiration. For English, there are six forms of articulation [20]: stops, nasal stops, fricatives, approximants, affricatives, and flaps. Consonants in English are further described by three parameters [21]: place of articulation, manner of articulation, and voicing. All of these articulators and parameters have been successfully modeled and can be identified by acoustic means with extremely high accuracy in quiet environments. However, in noise, or with speaker stress, the accuracy of these phoneme-based models fails.

As in previous model examples, the size of the phoneme provides little room for error in detection and identification. Thus researchers began to argue that the community was using a too small a unit of information. But what was available to the community that was in between a syllable and an allophone? The answer is our current state of the art unit of speech.

3.2.4 Diphones and Triphones

The current state of the practice is to use one of two blending technique of allophones known a diphones or triphones. In an effort to avoid interallophonic effects, researchers developed diphones. A diphone is the second-half of one allophone concatenated with the first-half of another allophone [22]. The center of a diphone is the boundary of the two allophones. This permits the interallophonic phenomena to be modeled. The number of reference models climbs. One hundred allophones produces 4,950 diphones if all combinations are used, which they are not. However, approximately 5000 reference models are manageable with modern processing equipment and this model has no exceptions and essential covers all the words in English.

Triphones offer the modeling of the interallophonic effects like diphones, but are large enough to reveal contextual information supported by a vocabulary. As the names implies, triphones use three phonemes. The center is relatively invariant with respect to a fully modeled diphone reference model which is present. The endpoints vary with context and are the center of the modeling effort to capture the articulation effects. The number of triphone models needed for a practical system is enormous. They are so numerous that for practical applications they are collected and modeled automatically from large speech databases. To the extent possible, a wide variety of speakers and dialects are desired for a high degree of speaker-independence. To help limit the scope of the collection and modeling effort, if possible, the vocabulary to support the system is identified. Thus the triphone collection and modeling may be limited to support the vocabulary helping limit the size of the model. A speaker adaptation algorithm is required to adjust the reference models of the triphones to the individual speaker. For large vocabularies it can take weeks or even months to complete the adaptation process. These systems are intended for long use by the same speakers.

The current state of the art consists of highly optimized diphone, triphone and even experimental N-phone models. While providing better modeling flexibility than word(s) and syllable-based models in medium or large vocabulary applications, so far none of these models have proven to be successful in large vocabulary noisy environments or effect in the presence of speaker stress. These models do not overcome speaker variation, dialect variation, and the articulatory degradation that normally accompanies continuously spoken speech. Speech extraction continues to represent a hard unsolved problem to the ASR community.

3.3 Latency

Similar to noise, latency is aggravated by the number of channels and processes encountered throughout the system as it executes. The major sources of latency include: communications channels, process transfer channels; internal or external process holds, distributed or internal data accesses, processing performance, and system support bus sizes. A typical commercial quality ASR application when fully trained and operated in a noise environment of 10db SNR or better can be expected to deliver response in 500 to 3000 milliseconds. In higher noise environments or when using a speaker independent system-latency can be as great as 10 seconds. This latter delay is unacceptable in most commercial applications, but is specifically unacceptable in military applications. Latency in cumulative and any decreases along the system path yield performance enhancement. Here are some techniques being researched to help with latency for ASR.

3.3.1 Increased Computational Power

Advances in this technology are out of the direct hands of the ASR community in so much as CPU and bus performance is concerned; however, like other applications we gain performance with every sub-system advancement in the computer. In addition to simple clock-speed enhancements and bus size, new designs support parallel processing which help reduce the latency created by data transfer between linear processes. These reductions in latency are benefits we get without direct improvements in the ASR application. More reductions are available in areas of the computer unique to the audio requirements of ASR and other acoustic-based applications.

3.3.2 Preprocessing Speech and Noise

New co-processors supporting audio input are being refined by members of the acoustic and ASR communities [23]. Digitized audio data files are large, even when compressed and the processes to evaluated audio in noisy or multi-speaker environments require enormous computational support. Continued development of dedicated audio processing components with floating point calculation routines will achieve reductions in latency that are not attributed directly to the ASR system's internal processes. At this point in the development of portable computing, most hand held PDA devices use processors that only support integer mathematics.

This requires writing computation routines in software rather than system calls to floating point calculations. This requires high quality programmers dedicated to elegant design for efficiency and speed in all calculations. Most research developers are adding separate audio processing capability or sophisticated pre-processing microphones to enhance the performance levels for their PDA-based prototypes. Speech and noise processing has good funding beyond the ASR community and the potential for improvement in noisy environments is good.

3.3.3 System Process Engineering

The final critical issue of latency is integration of the ASR system. This integration has two aspects: intra-integration of the ASR system with its own internal processes, and an external integration with the systems that the ASR will support. The current team of researchers and developers are working of functional process improvement to ensure minimal latency within the ASR system. However, the community has not put a major effort into the performance parameters required to fully integrate into a functional C^2 system with ASR [24]. Latency will not be dependent only on the response time of the core C^2 application and the ASR, but a calculation of the entire system of systems design model for future C^2 systems. Assuming the ASR system will not be a distributed network design, the latency impacts can be estimated with a fair accuracy. If the ASR is distributed (a possible solution for far forward deployment in computational stressed environments), latency will need to be monitored by some Quality of Service (QOS) manager within the ISO computer stack [25]. Regardless of the scenario of employment, the integration component of this effort for external integration is immature. All current and future ASR projects, as well as their C^2 counterparts need to look at latency with an eye on user satisfaction and mission requirements.

3.4 Error Rate

The issue of error rate is aggravated in ASR systems due to the modular makeup of the systems design. All current state of the art ASR systems are component designed. The audio input, speech extraction, quantization, and other components are all capable of introducing small amounts of error that are passed on from process to process. Unlike other information systems that can have trusted or verified data, the speech (as impacted by noise, stress, or dialect), can be easily misinterpreted at any stage of the ASR process and the error magnified by subsequent processes. Certain ASR paradigms are more subject to these cascading error effects than others.

For example, the use of a universal symbolic language to represent meaning in a language neutral design creates two additional error points as the systems translates to and from the inter-lingua process. Whereas single language (e.g., English) or direct language to language (e.g., English to Spanish), have only a single translation point of error. To counter accumulating errors, many systems introduce self correcting algorithms or direct human verification methods at strategic points in the process. An example was the creation of the language model to validate that the string of words being generated represented possible intelligent speech. Now that systems are being used in harsh environments, accumulative errors are more critical, and research is starting to develop more novel approaches to reducing error without human intervention. Within the research realms, resolving the cascading error effect of component based speech systems may well prove one of the most significant areas to support ASR in tactical environments. The goal will be to not only eliminate the errors, but also to do it with adding any significant latency.

4.0 Expectations

Unrealistic expectations were a major contributor to the failures of earlier attempts to introduce ASR into military and commercial systems. Three communities will be briefly covered in this section. The requirements, user, and developers all must have realistic expectations and driven by achievable requirements if a new generation of ASR technology is to be integrated into next generation C^2 systems.

4.1 Requirements

There have been several rapid prototype projects exploiting and evaluating ASR technology in field environments: *Phraselator* and *Babylon* projects from DARPA and the *Language and Speech Exploitation Recourses* (LASER) ACTD [26]. However these demonstration and proof of concept type efforts never created a stable set of requirements. That effort has begun with the Sequoyah program [27]. Sequoyah is to be the first program of record for development of multilingual translation and ASR technologies for integration into Army and Joint systems. Sequoyah is currently being stood up at PEO EIWS and is sponsored by the Combatant Command Interoperability Program Office located at Fort Monmouth, NJ. It is here that the initial draft requirements, based on the LASER ACTD and the DARPA projects, will be generated. The first step is the development of the Initial Capabilities Document, the ICD. As of

the publication of this paper, the ICD is being drafted and is expected to be staffed by 3rd quarter FY 2006 [28]. The development of a realistically achievable set of performance parameters and key performance parameters are essential to allowing the development community to build a practical ASR system. It is incumbent on this community to evaluate what features can and cannot be exploited in the initial delivery from our developers.

4.2 Users

Several Commercial Off The Shelf (COTS) and semi-militarized products have been delivered to operational forces in support of Operation Enduring Freedom (EOF) and Operation Iraqi Freedom (OIF). Most of these were developed by the Army Research Laboratory (ARL) and the Defense Research Projects Agency (DARPA). The most successful (as determined by most produced and requested by operators) has been the DARPA developed Phraselator [29]. Developed under a Small Business Innovative Research (SBIR) effort, by the submission of this paper over 1000 have been produced and delivered to the field. The Phraselator has gone through four major version enhancements all based on user needs to support field operations. Based on these lessons learned experiences, here is a list of major user requirements and their expectations for field use. These parameters (not exhaustive) should be carefully considered along with the formal ICD's technology gaps and requirements when developing and packaging a system for delivery in an operational environment. The final prototypes should have extensive user tests to ensure performance levels in operational environments support the user's mission. If not, the technology will be rebuffed by the user and another generation of users will be tainted against the insertion of ASR technology.

4.3 Developer

At this stage of capabilities development, the software requirements are critical mainly from the perspective of achieving easy system-of-systems integration. Beyond integration, there are three additional factors that require careful monitoring as the system is matured: power, weight, and packaging. Disregarding any of these factors could destroy developer expectations by created system requirements beyond current or next generation hardware and power capabilities.

4.3.1 Power

The system needs to have its own power source (batteries) capable of supporting an operation of 10 days without recharging. Through experience and observation, the DARPA program office determined that to meet this need, the battery must have charge holding capability of 12 days with support for 10 hours of "on" time within that 12 days [30]. Here "on" is defined as fully functional with no sleep system degradation of the CPU. With sleep support, days and hours of operation may be extended beyond the minimum requirement. The system should support charging from any power source. The latest DARPA version uses a switching power supply technology capable of charging the system from any voltage (3-220 VAC or VDC) at any frequency (40, through 100 Hertz (AC)) or DC. This was special important for overseas and multination operations. The final design request was that the rechargeable battery pack should be removable and replaceable with conventional "AA" batteries. These "AA" batteries in a soldier's environment are common and could be used in an emergency.

4.3.2 Weight and Packaging

Whether the user was a special operations team or a conventional force, everyone expected the system to be rugged and light weight. The first version of the Phraselator was a notebook computer with an integrated noise canceling microphone/speaker package. It became immediately clear that users needed and integrated uni-body package [31]. The second version was developed on an iPAQ; however, the integrated microphone and computational power proved unacceptable. The current solution, the Phraselator 2000 is a semi-militarized package providing a single package with sufficient battery life and computational power to run the ASR package. The weight of the Phraselator 2000 is 16 ounces – this weight seems to be the maximum for special operators and a preferred weight for conventional forces. Any future system should not exceed these weight and package parameters.

5.0 Risk

5.1 Program and Product Risks

Programmatic risk for a technology insertion is best assessed by the Technology Readiness Level (TRL) used by the Department of Defense [32]. The evaluation of TRL is for this component-based technology is categorized on the demands of the environment (noise and

stress levels) required for successful operation of the supporting C^2 system. Here is an assessment based on interviews with research developers, requirements officials, and the current DARPA program manager [33]:

1. For "office environments" similar to standard indoor and non-industrial commercial business situations, the current set of commercial ASR products is assessed to be TRL level 8: *Actual system completed and 'flight qualified' through test and demonstration*. The product will only need minor modification to operate with selected office equipment (or systems) and training for adaptation to selected noise reduction microphones used in the office environment.
2. For Tactical Operations Center (TOC) and other field operations where reliable power is available and major C^2 systems are operational, commercial ASR systems are not advised. For noise robust industrial/military ASR systems under development, the technology maturity is assessed to be TRL level 5: *Component and /or breadboard validation in a relevant environment*. Substantial development is still required to overcome the noise and speech stresses experienced in a TOC. Use at this point is recommended only for demonstration and concept validation or as part of a forward deployed experimental command center for evaluation by a voluntary, fully-informed field commander.
3. For forward operations in a mounted or dismounted environment in conjunction with far-forward C^2 or data collection/situational awareness systems, ASR technology will be challenged with noise (both stationary and non-stationary) and extreme variations in user speech stress. This environment is harsh, power constrained, and likely to experience hostile fire. Very few commercial based systems have been used in this environment. Success has been experienced only for applications with very small vocabularies. The most successful application has been for hand-held translation devices such as the *Phraselator* used in support of Special Operations Forces (SOF) in OIF and OEF. Further use of commercial systems in this environment is not recommended. For the noise robust industrial/military ASR systems under development, the technology maturity is assessed to be TRL level 4: *Component and/or breadboard validation in laboratory environment*. The greatest risk here is in maintaining the system

effectiveness and accuracy experienced in the larger power and computationally hungry systems designed for use in the TOC, but applied to the forward operating environment with their limited platforms and weight restrictions. At this stage of development, only highly customized prototype systems given to enthusiastic volunteer evaluators is recommended.

5.2 System Integration Risk

Integration risks are most likely to manifest themselves as system latency or data error input through incorrect ASR translations. Commercial ASR systems used in office environments have substantial error correcting capability. When fully trained and used on modern computers, they have little latency. Thus for office environment, use of ASR technology offers minimal integration risk [34].

Integration risk for TOC-oriented or forward-oriented C^2 systems is very high [35]. The TRL levels are still too low fully assess the level of effort required to ensure proper operation of the base C^2 system once the ASR technology is inserted. For the immediate future, it is recommended that only stand-alone systems or screen navigation systems use ASR at the TOC and forward.

6.0 Conclusion

Given the dynamics of the operational environment and the relative immaturity of noise robust ASR systems, the technology at this time is only deployable in systems requiring very small vocabularies (1-500 words) for use in speech translation systems, as a direct control, or system navigation and retrieval functions. These systems will most assuredly be speaker dependent (specific user voice training). For systems with speaker independent (no user training) applications in noise controlled office environments, systems requiring medium (2500 words) to large (5000+ words) vocabularies are viable. Based on the current R&D work and an anticipated increase in current funding over the next three years, the requirements and program communities should plan (Pre-Planned Product Improvement) on the insertion of robust ASR systems with large vocabularies by the 2011 timeframe [36]. Program offices should anticipate complete redesign of microphone and speaker systems to support the noise robust functionality during this same timeframe. Computational power requirements will also need to be carefully monitored if a dedicated CPU for ASR is not available. Noise robust ASR is becoming a reality,

this is not the time to suspend or delay research dollars. The commercial community will not develop noise robust technologies on there own [37]. The expected market is just too small outside the military community to warrant commercial development. Thus it is up to organizations like DARPA, ARL, NRL, and AFRL to continue to push the envelope on noise and speaker stress technology. The core research is complete; all we need now are the developmental dollars to get the components integrated and the first prototype in the field.
WORD COUNT: 7445.

Endnotes

1 The observation made in 1965 by Gordon E. Moore, co-founder of Intel, that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. In subsequent years, the pace slowed down a bit, but data density has doubled approximately every 18 months, and this is the current definition of Moore's Law, which Moore himself has blessed. Most experts, including Moore himself, expect Moore's Law to hold for at least another two decades. Available at: http://www.webopedia.com/TERM/M/Moores_Law.html (23 March 2006).

2 Ibid.

3 James D. Bass. *Breaking the Local Optima Paradigm: DARPA Speech Research Initiatives in Multi-Modal and Other Technologies.* Proceedings of the Automatic Speech Recognition: Challenges for the new millennium workshop. Paris, France, 18-20 September, 2000, pages 241-243.

4 R. Rosenfeld. *Two decades of Statistical Language Modeling: Where Do We Go From Here?* Proceedings from the Workshop on Spoken Language and Understanding. Summit, NJ, 6-9 February 2000.

5 B.H. Juang and S. Furui. *Progress & Challenges in Automatic Recognition and Understanding of Spoken Language.* Proceedings from the Workshop on Spoken Language and Understanding. Summit, NJ, 6-9 February 2000.

6 R. Rodman. *Computer Speech Technology*, (Boston, MA: Artech House, 1999), 51-93.

7 R. Rosenfeld. *Two decades of Statistical Language Modeling: Where Do We Go From Here?*.

8 R. Rodman. *Computer Speech Technology*, 113-114.

9 Filipp Korkmazskiy, Frank Soong, and Olivier Siohan. *Constrained Spectrum Normalization for Robust Speech Recognition in Noise*. Proceedings of the Automatic Speech Recognition: Challenges for the new millennium workshop. Paris, France, 18-20 September, 2000, pages 58-63.

10 M.L. Seltzer, B. Raj, and R.M. Stern. *Speech recognizer-based microphone array processing for robust hands-free speech recognition*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 13-17 May 2002, pages I-897 – I-900.

11 Sangita Sharma, Dan Ellis, Sachin Kajarekar, Pratibha Jain and Hyneck Hermansky. *Feature Extraction Using Non-linear Transformation for Robust Speech Recognition on the Aurora Database*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5-9 June 2000, pages 1117-1120.

12 Olivier Siohan, Tor André Myrvoll and Chin-Hui Lee. *Structural Maximum A Posteriori Linear Regression for Fast HMM Adaptation*. Proceedings of the Automatic Speech Recognition: Challenges for the new millennium workshop. Paris, France, 18-20 September, 2000, 120-127.

13 R. Rosenfeld. *Two decades of Statistical Language Modeling: Where Do We Go From Here?*.

14 Ace J. Sarich, "Phraselator," http://www.phraselator.com/ (23 March 2006).

15 R. Rodman. *Computer Speech Technology,* 229.

16 Ibid.

17 "Cochlear Filter Bank with Switched-Capacitor Circuits", http://www.isr.umd.edu/ISR/accomplishments/009_CochlearFilterBank/.

18 R. Rodman. *Computer Speech Technology*, 62-69.

19 Ibid.

20 L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*, (Englewood Cliffs, NJ, Prentice Hall, 1993), 20-42.

21 R. Rodman. *Computer Speech Technology*, 6-9.

22 R. Rodman. *Computer Speech Technology*, 120-121.

23 M.L. Seltzer, B. Raj, and R.M. Stern. *Speech recognizer-based microphone array processing for robust hands-free speech recognition*.

24 Robert Noelsch, Commander, USN (Combatant Command Interoperability Program Office), in discussion with the author, December 2005.

25 Ibid.

26 Mari Maeda (Defense Advanced Research Projects Agency), in discussion with the author, December 2005.

27 Ibid.

28 As of publication of this paper, a draft ICD is available from the following source:

29 Mari Maeda (Defense Advanced Research Projects Agency), in discussion with the author, December 2005.

30 Ibid.

31 Ibid.

32 "Technology Readiness Level (TRL), DoD 5000.2-R Appendix A6-4", http://www.onr.navy.mil/fncs/explog/explog/docs/overview/trldefinitions.pdf#search='technology%20readiness%20level' (23 March 2006).

33 Mari Maeda (Defense Advanced Research Projects Agency), in discussion with the author, December 2005.

34 Ibid.

35 Ibid.

36 Robert Noelsch, Commander, USN (Combatant Command Interoperability Program Office), in discussion with the author, December 2005.

37 Ibid.